# Modelling built-settlements between remotely-sensed observations

Jeremiah J. Nieves, Alessandro Sorichetta, Catherine Linard, Maksym Bondarenko, Jessica E. Steele, Forrest R. Stevens, Andrea E. Gaughan, Alessandra Carioli, Donna J. Clarke, Thomas Esch, & Andrew J. Tatem

## ABSTRACT

Mapping settlement extents at the annual time step has a wide variety of applications in demography, public health, sustainable development, and many other fields. Recently, while more multitemporal urban feature or human settlement datasets have become available, issues still exist in remotely-sensed imagery due to coverage, adverse atmospheric conditions, and expenses involved in producing such feature sets. These challenges make it difficult to increase temporal coverage while maintaining high fidelity in the spatial resolution. Here we demonstrate an interpolative and flexible modeling framework for producing annual built-settlement extents. We use a combined technique of random forest and spatio-temporal dasymetric modeling with open source subnational data to produce annual 100m x 100m resolution binary settlement maps in four test countries of varying environmental and developmental contexts for test periods of five-year gaps. We find that in the majority of years, across all study areas, the model correctly identified between 85-99% of pixels that transition to built-settlement. Additionally, with few exceptions, the model substantially out performed a model that gave every pixel equal chance of transitioning to the category "built" in each year. This modelling framework shows strong promise for filling gaps in cross-sectional urban feature datasets derived from remotely-sensed imagery, provide a base upon which to create future built/settlement extent projections, and further explore the relationships between built area and population dynamics.

**Keywords**
Built, urban growth, random forest, dasymetric, population,

## ACKNOWLEDGEMENTS

## AUTHOR CONTRIBUTIONS

JJN, AS, JES, and AJT designed research. FRS, AEG, CL and JJN contributed to previous model concepts that resulted in the presented model realization. DC and AC contributed significant knowledge transfer on bootstrapping and growth curves. JJN carried out analyses and research. JJN, MB, and TE provided data and/or carried out additional data pre-processing. JJN wrote the modelling script with MB providing the code framework for the global data production. JJN wrote the initial manuscript draft with contributions and edits from all other authors.
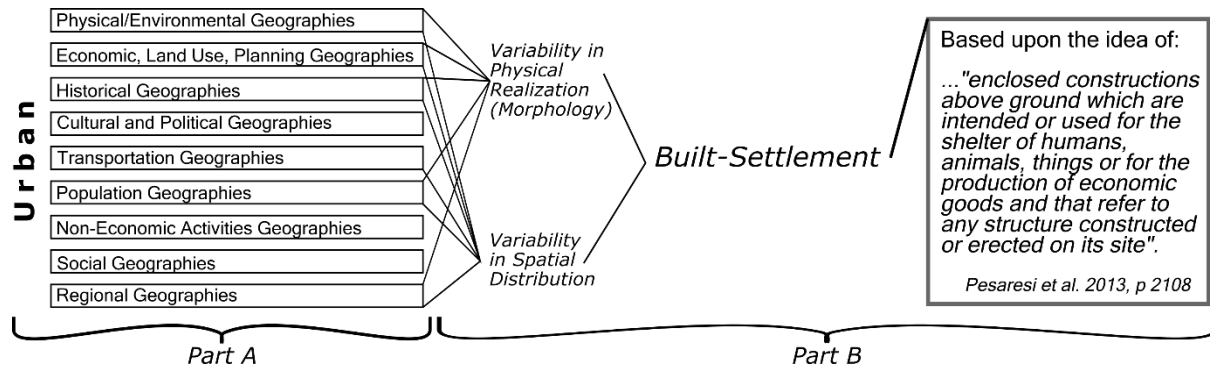
## 1. INTRODUCTION

Having time series of regular and consistent observations of built settlement extents is important given that forecasted growth of populations within dense urban areas are expected to continue through 2050, with much of that increase will occur within Africa and Asia (1,2). Further, rapidly changing magnitudes and distributions of both settlements and populations have significant implications for sustainability (3), climate change (4,5), and public health (6,7), amongst others. At local and regional levels, the availability (or non-availability) and accuracy of built-settlement extent data affect measured population distributions, densities, and built type (e.g. urban, peri-urban, and rural) used to inform and shape policies. The 2030 Agenda for Sustainable Development, which have a focus on accounting for and including "all people everywhere", reinforced the need for readily and globally available baseline data to guide efforts and measure progress toward its Sustainable Development Goals (SDGs) (8).

Urban has been defined in many ways across many fields with different definitions existing even within the same field depending upon the specific application. Many countries define urban as a function of some population magnitude/density threshold or based upon administrative jurisdictions and functional economic areas and activities (9,10). While not conducive to applications requiring global consistency in definitions (11), none of these definitions of the concept of urban are objectively wrong. Urban, whose formal yet vague language definition is "of, relating to, characteristic of, or constituting a city" (12) is a complex amalgamation of the physical environment, population, economics, movements, and

connectivity (3,13–30). Figure 1, Part A gives a generalized diagrammatic view of the factors contributing to the concept of urban.

**Figure 1.**  Conceptual relations and definition of "built-settlement" (Part B) within the generalized concept of "urban" (Part A) and the broad, non-exhaustive contributing factors that make these concepts.



As a result, many studies have turned to a definition based upon the remotely-sensed physical features of urban areas, i.e. the built-environment. However, even reducing the definitional scope of urban to its physical dimension, the form of built-environment can widely vary across space and time due to materials used, differences in urban morphology, and the surrounding environmental context (18,31,32). Initially, remotely sensed urban definitions were optically-based thematic classifications of land cover, typically captured the "built-environment," including buildings, roads, runways, and, sometimes erroneously, bare soil (18,33–35). Later improvements using supporting information about the surrounding environment and vegetation during post-processing helped discern the true built-environment from the surrounding land covers (18). Other notable advances include the use of high resolution orthographic imagery to detect subtle short-term built-environment change in China (36) and the use of Landsat imagery to create multi-temporal thematic representations of the built environment across the globe (37).

Coinciding with advances in imagery, statistical methods, and computational resource availability, high-resolution datasets with global extent have been created either through combining multi-source optical imagery with contrast detection methods (38,39) or utilizing

Synthetic Aperture Radar (SAR) data with object-based image analysis to refine the capture of urban features, with a focus on vertical human-made structures, while attempting exclude other anthropogenic land covers (40). However, it remains a challenge to produce a consistent global urban feature product while maintaining high temporal and spatial fidelity, meaning most of the global multi-temporal urban feature data sets refer to few time points across a larger time period. Further, the resource cost of producing these data remains relatively high (41) and there can be pre-existing gaps in the input data, due to selected sensor/platform characteristics or problems and adverse atmospheric conditions, prior to the other fidelity considerations. While there is now a global abundance of high-resolution imagery, with various instruments and revisit times on various platforms, not every image is suitable for extracting urban features due to the aforementioned reasons and/or because the processing cost is either not economical for the global production of high frequency urban feature extractions or the funding for such endeavours is not available.

One way to address these issues is to leverage years where remotely-sensed urban feature extractions with high spatial fidelity are available and interpolate for missing time points and areas of interest by modelling between the coverage of remotely-sensed urban feature data. Overall, urban feature/built-environment growth models have disproportionately focused on high-income countries, which have different dynamics than low- and middle-income countries (1,2,21,42), and most have been limited to city or regionally specific models (42–51). Previous methods of modelling urban feature/built-environment growth across space and time at the continental and global scales include land cover/land use transition models (52,53) and cellular automata models (50,54,55), with features or thematic classes extracted from remotely-sensed imagery being the primary source of cross-sectional input for these models (18,38–40,56). Readers are referred to Li & Gong (57) and Sante et al.

(50) for comprehensive reviews of the wide field of cellular automata models as applied to urban feature/built-environment growth modelling.

Of the few models predicting urban feature/built-environment growth across the globe within a standardized framework, almost none have explicit spatial prediction finer than country level summaries (20,21). Global models that did provide discrete spatial predictions, did not allow local sub-national variation to drive the modelled changes or had not been assessed against comparable existing datasets (20,42,58,59). Because the focus of this study is on modelling urban feature/built-environment extents that better represent where people may be located, we adopted the Global Human Settlement Layer (GHSL) concept of "built-settlement" (BS), which is defined as, "…enclosed constructions above ground which are intended for the shelter of humans, animals, things or for the production of economic goods and that refer to any structure constructed or erected on its site." (38). We generalized the definition of built-up structure to include other urban feature datasets that attempt to refine their thematic or feature set to represent buildings associated with human activities while attempting to exclude more general impervious surfaces, such as roads, parking lots, and runways. With the adopted definition of BS, the analogue to the process of "urbanization", is taken within a remote sensing context to be the physical transition from a non-BS area to a BS area.

With these definitions in mind, we leveraged recent advances in multi-temporal global BS feature datasets, global environmental datasets, subnational census-based population data, and computational methods to develop a globally applicable and flexible modelling framework. Our specific objectives were to i) determine if random forests can reasonably predict non-BS to BS transitions, ii) use the random forest to create an automated globally applicable framework that can produce spatially explicit estimates of BS extent at annual

6

steps using sub-nationally driven geospatial covariates and population counts, iii) validate the

model performance and compare the model predictions to the "observed" data.

## 2. METHODS AND DATA

### 2.1 Study Areas and Data

We selected four countries (Table 1) from across the globe to capture a variety of BS

morphologies, contexts, and evolutions as well as to demonstrate the flexibility of the model

for differing  spatial resolutions of input census-based population data, as measured by the

average spatial resolution (60). The countries selected here were Panama, Switzerland,

Uganda, and Vietnam, which exist in rather contrasting environmental/urban biomes (18) and

topographies and represent quite different cultural and developmental contexts (from low-,

middle-, to high-income.

**Table 1.** Summary of built-settlement transition data by country and period. Areal units here are pixels (~100m) as that is the unit handled by the model which looks at relative areal changes as opposed to absolute areal changes.

| Country | Average Spatial Resolution[a] | Period | Initial Non-Built Area (pixels) | Observed Transitions |
|---|---|---|---|---|
| Panama | 10.9 km | 2000-2005 | 8,901,004 | 0.03 % |
|  |  | 2005-2010 | 8,898,679 | 0.09 % |
|  |  | 2010-2015 | 8,890,339 | 0.75 % |
| Switzerland | 3.9 km | 2000-2005 | 6,816,510 | 1.56 % |
|  |  | 2005-2010 | 6,710,069 | 0.08 % |
|  |  | 2010-2015 | 6,704,973 | 0.01 % |
| Uganda | 12.2 km | 2000-2005 | 28,231,555 | 0.07 % |
|  |  | 2005-2010 | 28,210,425 | 0.04 % |
|  |  | 2010-2015 | 28,200,084 | 0.04 % |
| Vietnam | 21.7 km | 2000-2005 | 40,108,425 | 0.11 % |
|  |  | 2005-2010 | 40,063,545 | 0.18 % |
|  |  | 2010-2015 | 39,990,858 | 0.38 % |

a  Average spatial resolution is the square root of the average subnational area, in $km^2$, and can be thought of as analogous to pixel resolution with smaller values indicating finer areal data and vice versa (60)

We also selected a number of covariates based upon previous urban feature/built-

environment models (42,55,61) to give immediate environmental/land cover context and

connectivity of areas to urban/built agglomerations to the model. Ultimately, the model is not

dependent on any specific covariates, retaining a level of flexibility for use in a wide variety of applications. For example, we have also tested the model using a minimal set of globally available predictive covariates to provide input to other modelling efforts and avoid potential covariate conflict for end users. In the case presented here, covariates available for each year were either directly used in the model (i.e. esa_cls190 and DMSP night-time lights) or aggregated temporally (i.e., VIIRS night-time lights), while covariates available for just one point in time were used only if reasonably assumed to be time invariant (Table 2). All covariates were pre-processed, appropriately resampled to a common spatial grid having a resolution of 3 arc seconds; with the latter chosen as a compromise between the higher resolutions of some covariates (Table 2) and the coarser covariates such as the ESA land cover datasets. All data used to restrict the area of modelling and inform the redistribution of transitions are also detailed in Table 2. Further details on pre-processing of specific covariates are provided in the Supplemental Material.

**Table 2.** Data used for predicting the annual number of non-BS to BS transitions at the unit level (i.e. demand quantification), the pixel level probability surface of those transitions, and other data utilized in the spatial allocation procedures of the model.

| Covariate | Variable Name(s) in Random Forest | Description | Use[b] | Time Point(s) | Original Spatial Resolution(s) | Data Source(s) |
|---|---|---|---|---|---|---|
| Built-settlement[c] | esa_cls190 | Binary BS extents | Demand Quantification and Spatial Allocation | 2000 2005 2010 2015 | 10 arc sec | (62) |
| DTE Built-settlement | esa_cls190_dst_*<year>* | Distance to the nearest BS edge | Spatial Allocation | 2000 | 10 arc sec | (62) |
| Proportion Built-settlement 1,5,10,15 | esa_cls190_prp_*<radius>*_*<year>* | Proportion of pixels that are BS within 1,5,10, or 15 pixel radius | Spatial Allocation | 2000 | 10 arc sec | (62) |
| Elevation | Topo | Elevation of terrain | Spatial Allocation | 2000 – Time Invariant | 3 arc seconds | (63) |
| Slope | Slope | Slope of terrain | Spatial Allocation | 2000 – Time Invariant | 3 arc seconds | (63) |
| DTE Protected Areas Category 1 | wdpa_cat1_dst_2015 | Distance to the nearest level 1 protected area edge | Spatial Allocation | 2015 | Vector | (64) |
| Water | --- | Areas of water to restrict areas of model prediction | Restrictive Mask | | 5 arc second | (65) |
| Subnational Population | --- | Annual population by sub-national units | Demand Quantification | 2000 - 2020, annually | Vector | (66) |
| Weighted Lights-at-Night (LAN) | ---- | Annual lagged and sub-national unit normalised LAN | Spatial Allocation | 2000-2016, annually | 30 arc second (2000-2011) 15 arc second (2012-2016) | DMSP (67,68) VIIRS(68,69) |
| Travel Time 50k | tt50k | Travel time to the nearest city centre containing at least 50,000 people | Spatial Allocation | 2000 | 30 arc second | (70) |

| | | | | | | |
|---|---|---|---|---|---|---|
| Urban Accessibility 2015 | urbanaccessibility_2015 | Travel time to the nearest city edge | Spatial Allocation | 2015 | 30 arc second | (71) |
| ESA CCI Land Cover (LC) Class [a] | ccilc_dst*<class number>_<year>* | Distance to nearest edge of individual land cover classes | Spatial Allocation | 2000 | 10 arc second | (62) |
| Distance to OpenStreet Map (OSM) Rivers | osmriv_dst | Distance to nearest OSM river feature | Spatial Allocation | 2017 | Vector | (72) |
| Distance to OpenStreet Map (OSM) Roads | osmroa_dst | Distance to nearest OSM road feature | Spatial Allocation | 2017 | Vector | (72) |
| Average Precipitation | wclin_prec | Mean Precipitation | Spatial Allocation | 1950 - 2000 | 30 arc sec | (73) |
| Average Temperature | wclim_temp | Mean temperature | Spatial Allocation | 1950 - 2000 | 30 arc sec | (73) |

a Some classes were collapsed: 10-30 → 11; 40-120 → 40; 150-153 → 150; 160-180 → 160 (74)
b Covariates involved in Demand Quantification were used to determine the demand for non-BS to BS transitions at the subnational unit level for every given year. Covariates involved in Spatial Allocation were either used as predictive covariates in the random forest calculated probabilities of transition or as a post-random forest year specific weight on those probabilities and the spatial allocation of transitions within each given unit area. Covariates used as restrictive masks prevented transitions from being allocated to these areas.
c The binary BS data utilized 2000, 2005, 2010, and 2015 as observed points in the dasymetric modelling process, but only derived covariates for 2000 were utilized in the random forest as predictive covariates

### 2.1.1. Built-Settlement Data

We chose to use the "Urban" thematic class, class 190, from the annual ESA CCI land cover dataset (hereafter ESA) for our study. It was selected for its annual coverage from 1992 to 2015, allowing for the exclusion of years in the modelling process for validation of outputs. For our period of interest, 2000 to 2015, the ESA data creates annual 10 arc sec resolution (~300m at Equator) datasets by looking for thematic class changes from a baseline land cover map, produced using MERIS imagery, using 30 arc second (~1 km at the Equator) SPOT VGT imagery (1999-2013) and PROBA-V imagery (2014-2015) (75). Prior to 2004, detected changes are delineated at 30 arc second resolution. Starting in 2004, if there are changes detected, then the individual pixels of change detected at 30 arc second are further delineated using 10 arc second MERIS or PROBA-V imagery (75). To reduce false detections, changes must be observed over two years or more (75). Furthermore, the Global Human Settlement Layer (GHSL) (38,39) and Global Urban Footprint (GUF) (40) datasets are utilized in defining the extents of the ESA Urban class (75), which thus incorporate elements of two BS datasets within the larger built-environment context. While still undergoing full validation, initial validation efforts estimate the 2015 Urban class user and producer accuracies between 86-88 percent and 51-60 percent, respectively (75). We also test and validate a single year from an alpha version of forthcoming multi-temporal World

Settlement Footprint (WSF) dataset, known as WSF Evolution (41), and present the results in the Supplementary Material.

### 2.1.2 Population Data

Annual population estimates from 2000 to 2020 for subnational areas were provided by the Center for International Earth Science Information Network (CIESIN) in tabular format with unique IDs corresponding to a consistent grid containing the corresponding unique subnational unit IDs (66). Populations and the areas of subnational units are based upon the Gridded Population of the World, version 4 (GPWv4) and as such follow the methods detailed in Doxsey-Whitfield et al. (66) for the interpolation of population between years of official counts or estimates.

## 2.2 Built-Settlement Growth Model (BSGM)

### 2.2.1 Overview

Here we interpolated for every year between a set of timepoints, $T = \{t0, t_1, t_2, ..., t1\}$ where $t0$ is the initial observed timepoint, $t1$ is the final observed timepoint, and all other times $t$ are points lying between $t0$ and $t1$ for which we had observed BS extents. The time between any two observed time points $t$ is referred to as a period, $p$. Within this study, $t0$ is 2000, $t1$ is 2015, and we also have observed time points at 2005 and 2010, however the framework can handle any regularly spaced intra-period time-step if the input data corresponds. Therefore, in this study, for the ESA informed models we are carrying out modelling on three time periods, 2000-2005, 2005-2010, 2010-2015, for a total of 12 years interpolated. We generalize the process to determine the timing and number of transitions for each time step independently for each subnational unit, hereafter unit, as follows:

1. Create a population map for all $t_{observed}$ time points in $T$.

2. At all $t_{observed}$, for each unit, extract the time-specific population count within the time specific BS extents and derive the corresponding average BS population density.

3. On a unit-by-unit basis, interpolate the extracted BS population count between each $t_{observed}$ using piecewise-fit logistic growth curves and BS population density by fitting natural cubic splines across all $t_{observed}$.

4. Estimate expected unit, time-step-specific, total BS extent area, in number of pixels, and the expected number of transitions for that time-step based upon predicted unit-specific total BS population and BS population density.

5. For every unit, adjust the expected transitions by the sum of all expected transitions across the given period, e.g. 2000-2005, and divide by the total observed changes, e.g. 2005 BS extents minus 2000 BS extents. Repeat for all periods.

6. For each unit, use the adjusted predicted transitions from step five as relative weights within a given unit to dasymetrically redistribute observed transitions from the larger source period to smaller individual time-steps, i.e. years, while maintaining the original number of transitions when all time-steps are summed. Repeat for all periods.

To spatially disaggregate the predicted transitions, we first train a Random Forest (RF) model (76) to produce a continuous surface representing the probability of a given grid pixel transitioning from non-BS to BS between $t0$ and $t1$. For every time-step, processing each subnational unit independently, we utilized unit normalized lagged lights-at-night (LAN) data to annually adjust the base RF-derived transition probabilities. The assumption behind this being that areas that underwent the largest increase in brightness, relative to the rest of the unit in a single time-step, have a higher probability of transitioning and vice versa. From the pixels that were known to have transitioned, as indicated in the input BS data, we

selected pixels with the $n^{th}$ highest probabilities for transition, where $n$ was equal to the

number of pixels predicted to transition for that time-step. We then converted those pixels to

BS, recorded the new BS extents, and used those extents as the basis for the next time-step of

transitions. This resulted in a series of regularly spaced time-specific binary spatial

predictions of the BS extents in raster format. A full process diagram is shown in Figure 1.

All modelling and analyses were carried out using R 3.4.2 (77) on the IRIDIS 4 high-

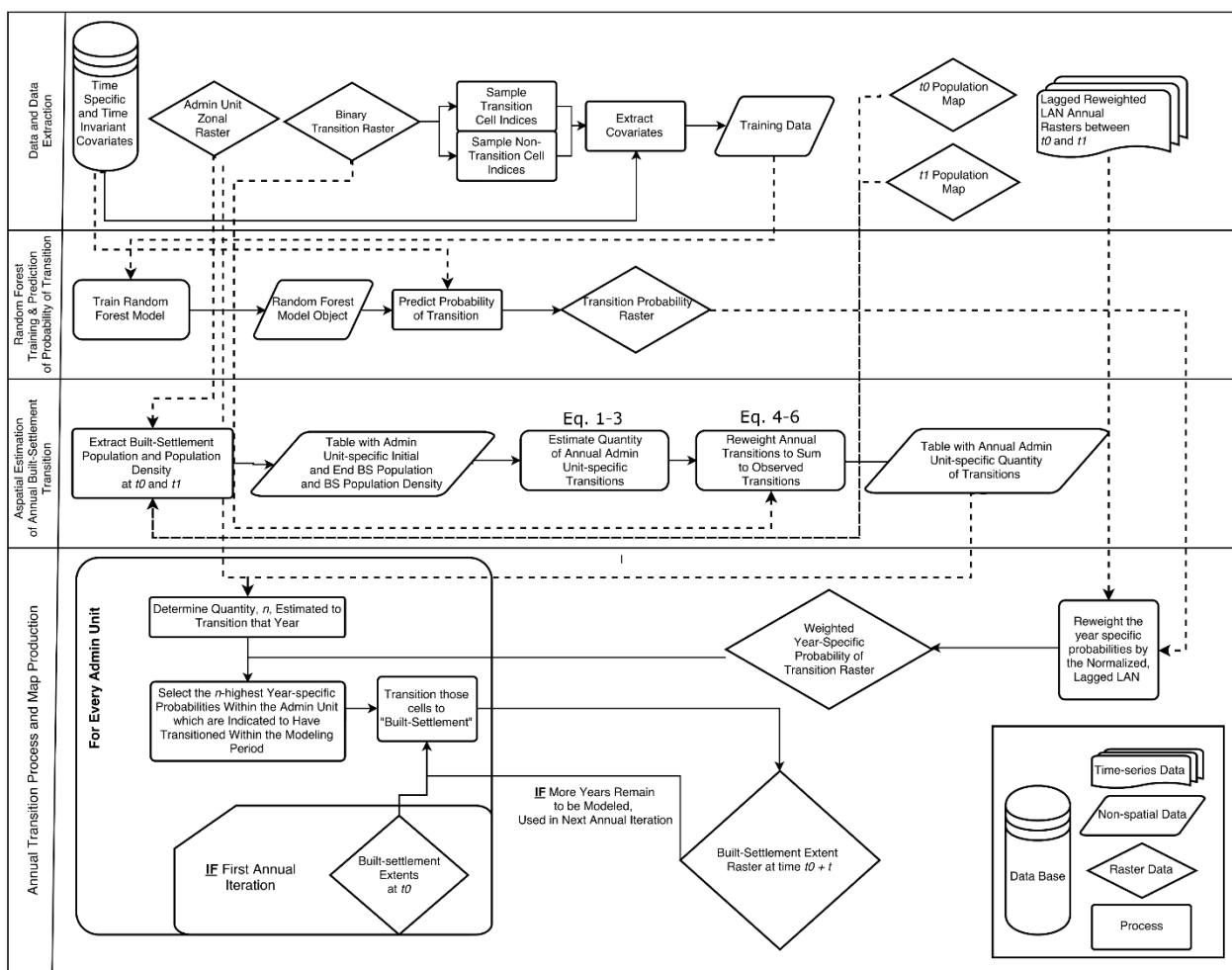performance computing cluster (see Supplemental Material for code).



**Figure 2.** Overview of the generalized modelling process for a case of only two observed timepoints, *t0* and *t1*, with references to utilized equations.

### 2.2.2 Random Forest (RF) Estimation of the Probability of Transition

Given that the binary dataset of transition/non-transition constitutes an intrinsic

"imbalanced set" (78), i.e. there are many more non-transitions than transitions; we adopted a

stratified random over/under-sampling method (78), similar to (42), as follows: (i) randomly sample 80 percent of the pixels that transitioned, up to 50,000 and, (ii) randomly sample an equal number of pixels that did not undergo transition. The choice for equal sampling of each stratum was determined by testing different relative proportions and samples sizes until finding the most consistent and best model results, balancing performance and efficiency.

To create a probability of transition surface for the complex and non-linear phenomenon of BS transition, we utilized a RF model to accurately and efficiently model across an entire country at the pixel level (i.e. 3 arc second spatial resolution). A RF model was selected for its robustness to noise, its automatability and efficiency, and its ability to capture non-linear and complex interactions (76). Further, in at least one study, RFs have been shown to perform equally, if not better, than other methods, such as support vector machines and logistic regression, when predicting the probability of transitioning from non-built-environment to built-environment (79).

We trained the classification RF on whether a pixel had transitioned between time *t0* and *t1* (1) or not transitioned (0) against the corresponding values of covariates at time *t0*. We used the constructed RF to predict for the entire given country. Rather than accepting the default output of the RF classifier, which outputs a single predicted class as indicated by the majority of the predictions of its individual constituent trees, we wanted a continuous, 0.00 to 1.00, probability of transitioning in to discriminate between high and low probabilities. Given that we trained the RF as a binary classifier, we took the mean of the individual tree predictions for each pixel. This class probability has a value between 0.00 and 1.00 represents the posterior probability of a pixel being classified by the RF as transitioning between *t0* and *t1* (76).

13

### 2.2.3 Population Mapping of Endpoints

To interpolate, we first needed a spatially-explicit best estimate of the subnational unit specific BS population at all observed timepoints in the modelling period. To get this we created a population surface using the available time-specific and, assumed, time-invariant covariates (see Supplemental Material, Table A2) using the WorldPop RF method, described in Stevens et al. (80), to dasymetrically redistribute the time-specific population totals from the subnational unit level to 3 arc second grid pixels (80–82).

For any given time point in the population modelling, we included the distance to nearest BS edge for the *t0* timepoint, i.e. 2000, as population relates to older parts of a BS agglomeration differently from newer ones (83). For example, if we were to model the population map of 2010 we would include the distance to nearest observed BS edge for 2010 as one of the predictive covariates as well as the distance to nearest BS edge corresponding to the observed 2000 BS extents. This was done to avoid centres of agglomerations being assigned artificially low population densities relative to the preceding modelled time point (83). For each subnational unit, we then extracted and summed the total populations that were spatially coincident with the BS extents and derived the corresponding BS population density for use in the BSGM predictive phases.

### 2.2.4 Transition Magnitude Estimation

To estimate the number of transitions for each time-step within the study period, we used the predicted BS population changes and the predicted changes in BS population density for every subnational unit. We first interpolated the BS population count of each unit $i$ for every year, BSPOP(t), by fitting logistic growth curves (84), in a piecewise manner, i.e. for each time-period between observed points, using the year-specific total population, $K_i(t)$, as the varying carrying capacity (85) as shown in Equation 1

$$BSPOP_i(t) = K_i(t) * \frac{e^{r_i * t + C_i}}{1 + e^{r_i * t + C_i}}$$  [Eq. 1]

where $t$ is the number of time steps from the given period's origin, e.g. for period 2000-2005

2002 corresponds to $t = 2$, and $r_i$ and $C_i$ are determined by fitting a least-squares linear

regression to the set of observed values after having been transformed via Equation 2.

$$\ln \left( \frac{BSPOP_{it_{observed}}}{K_{it_{observed}} - BSPOP_{it_{observed}}} \right)$$  [Eq. 2]

Logistic growth curves are widely used and accepted for modelling populations within

demography, ecology, and urban modelling (28,86–89). Batty (54) summarized, "Constrained

population growth reflecting both exponential change and capacity which, in turn, reflect

densities and congestion are simulated using various kinds of logistic growth." because, as

Sibly and Hone (90) note, "While environmental stressors have negative effects on

population growth rate, the same is true of population density, the case of negative linear

effects corresponding to the well-known logistic equation." as put forth first by Verhult (91)

in 1838. Furthermore, Ledent (28) showed that urbanization, the process of *population*

becoming urban, across time can be adequately summarized by "S-shaped curves",

specifically the functional logistic form. We followed this same underlying conceptual logic

using the above logistic curve with a dynamic limiting factor (Equation 1), i.e. the total

population of an area is the theoretical limit of the temporally coincident BS population

count.

Then, for each subnational unit, we interpolated  its average BS population density

across all observed time points in the study period using natural cubic splines (92) and the

observed points as the knots, i.e. for ESA 2000, 2005, 2010, and 2015. Because there is a lack

of literature and data on the actual or theoretical limits of human population density, we

selected cubic splines, which have a long history in demographic interpolation, including

interpolation of rates associated with urbanization processes (28,92). We are assuming that

the trend of population density change of a given subnational unit is smooth and continuous

across time, because short of some drastic (and unlikely and largely unaccountable

"population shocks" such as wars or natural catastrophes) event, at the annual time scale we

would not expect to see "cliffs" of population density change. With a piecewise linear

interpolation, we would see such sharp turns in rates of change. With our priority being that

the fit curve would match values of our observations, the rationale behind non-parametric

smoothing is to employ un-parametrized functions (i.e. splines) that adapt to the data, rather

than adapting the data to a given distribution, (i.e. a parametric approach). We did not choose

polynomial interpolation or a higher degree spline to avoid Runge's Phenomena where, as

data is added, the derivatives at each data point increases, resulting in large oscillations of

rates of change between data points.

Using the interpolated subnational unit BS population and the interpolated subnational

unit-average BS population density, we simply relate population and population density to

get the expected number of transitions in Equation 3

$$BSCNT_i(t) = \frac{BSPOP_i(t)}{BSD_i(t)} \qquad\qquad \text{[Eq. 3]}$$

where $BSD_i(t)$ is the unit, $i$, average BS population density at time $t$. See Supplemental

Materials for how predicted "negative growth" resulting from Equations 1-3 was handled.

### 2.2.5 Dasymetric Redistribution of Transitions Across Time

The number of predicted transitions derived from Equations 1-3 are not inherently

constrained by the observed transitions between any two observed time points. To match the

total number of observed transitions for the modelled period, we reweighted the transitions of

each time-step on a unit-by-unit basis. This is essentially a temporal dasymetric redistribution

of transitions from the larger source period, e.g. 2000-2005, to the smaller target periods of *t*,

e.g. 2001,2002, etc., based upon temporal information contained in the time-specific unit

population totals. To calculate the weight for each time-step, $w_t$, we write the calculation in

Equation 4 as:

$$w_{t_i} = \frac{B\widehat{SCNT}_i(t)}{\sum_1^k B\widehat{SCNT}_i(t)} \qquad \text{[Eq. 4]}$$

where $t$ is again relative to the given period from 1 to the last year $k$ and all $w_t$ for a given unit

$i$ sum to one. To obtain the temporally weighted transitions, $B\widehat{SCNT}_{it}$, we multiplied the

weight of each year by the observed number of transitions in Equation 5, rounding to the

nearest whole number for each year (see Supplemental Materials, section A4 for obtaining

agreement with rounding differences).

$$B\widehat{SCNT}_{ti} = round(w_{ti} * \Delta BSCNT_i) \qquad \text{[Eq. 5]}$$

Where $\Delta BSCNT_i$ is the number of observed transitions, in pixels, from non-BS to BS for a

given unit $i$. This allows the model to maintain agreement of transitions between the points

that we interpolated.


### *2.2.6 Spatially Disaggregating Transitions Using Annual Unit-specific LAN-weights*

For each period, we then processed the tabular predicted transitions into time-specific

BS extent maps, i.e. spatially allocated the transitions within each subnational unit for each

time-step. We spatially assigned the transitions within each unit using the RF-derived

transition probability surface adjusted by time-specific weights in the form of subnational

unit normalized LAN brightness differences, i.e. one time-step lags (see Supplemental

Material). Given that the non-BS to BS transition process is iterative in nature, we began by

taking the extents of the previous time-step, or the previous observed extents if $t$ was equal to

one. We limited the location(s) where transitions could be allocated within the subnational

unit to pixels where transitions were observed to have occurred, as defined by the input BS

data. For every one of those locations, $j$, assuming they weren't transitioned in previous steps,

we retrieved the transition probability as calculated in the transition probability surface. We

took this base probability of transition for every pixel *j* and adjusted it by the spatially

coincident lagged and weighted LAN data, for time-step *t*, using Equation 6:

$$P_{adj}(transition)_{ijt} = wLAN_{ijt} * P(transition)_{ij} \qquad\qquad \text{[Eq. 6]}$$

where $P(transition)_{ij}$ is the RF-derived transition probability where transition was

observed, $wLAN_{ijt}$ is the corresponding adjusted LAN difference observed for the time-step

*t,* and $P_{adj}(transition)_{ijt}$ is the adjusted probability of transition for the time-step *t* in a

given pixel *j* within the administrative unit *i*. Similar to previous models (42,53), we assumed

pixels with a higher probability of transition are more likely to transition before pixels with

lower probabilities. We selected the $n^{\text{th}}$ highest probabilities from the subset of potential

transition pixels, where *n* was equal to $\widehat{BSCNT}_{ti}$, changed the value of those selected pixels to

represent a transition to BS, and output the union of the new transitions and previous BS

extents as the predicted BS extents for that time-step. We repeated this procedure using the

newly produced extents for the preceding time-step as the base BS extent for the next time-

step's transition procedure, until all time-steps for the period were processed and then the

entire procedure was repeated until all periods had been processed.

### 2.3 Analyses

#### 2.3.1 Validation and Comparison Metrics

While the RF produces its own validation estimates (76), we tested the accuracy of

the RF classifier by randomly sampling 100,000 pixels, not utilized in the training of the RF,

for validation. We selected this sample size as we were able to obtain sample prevalence rates

equal to the known true prevalence rates of each country while still maintaining efficiency.

Based on this sample, we plotted receiver operator curves (ROCs) and, given the imbalanced

data (93,94)**,** precision recall curves (PRCs) with simulated perfect and random classifier

curves for comparison.

Here, for the ESA models, we are comparing the predicted extents to all withheld extents between 2000 and 2015. For every year of prediction, we determined whether a pixel was True Positive, False Positive, False Negative, or True Negative, TP, FP, FN, TN, respectively. Pixels used for validation of the modeled outcomes were limited only to pixels having been observed transitioning from non-BS to BS between the modeled periods for two related reasons:

1) Being an interpolative model, we constrained the areas of possible transition predictions to only the areas of observed transition. This limited the spatial uncertainty of the model between 2000 to 2005, 2005 to 2010, and 2010 to 2015 to no worse than the input data, although temporal uncertainty and the spatial configuration of transitions for any specific year between those periods remained. Essentially the same limiting of uncertainty seen with "volume preserving" spatial only dasymetric mapping, but extended to the temporal component as well as indicated in Mennis and Hultgren (82).

2) Given that we masked our predictions to only pixels we knew transitioned, if we were to have included pixels that we knew not to have transitioned, we would have grossly and erroneously inflated the error metrics as they would have had no chance of disagreeing.

Using the validation data for comparison, we calculated contingency table-based metrics to evaluate classification accuracy based primarily on the $F_1$ score (Table 3) which is the harmonic mean of recall and precision, the quantity disagreement (95), and the allocation disagreement (96). We aggregated the pixel level results, to the unit level and calculated the same metrics since precision, and by extension $F_1$, is sensitive to the corresponding prevalence and is subject to the modifiable areal unit problem (MAUP) (97).The MAUP not only reduces variance in value distributions the more the data are aggregated from their

original resolution (97), but will result in different prevalences with different subnational

area, i.e. zonal, configurations. The equations of the metrics calculated are listed in Table 3.

**Table 3.** Classification agreement metrics utilized and a brief description of each. The F1-score is interpreted as the harmonic mean of precision and recall. TP is "True Positive", FP is "False Positive", FN is "False Negative", and TN is "True Negative."

| Metric | Equation | Range and Interpretation |
|---|---|---|
| Recall (Sensitivity) (98) | $\dfrac{TP}{TP + FN}$ | 0 (no recall) – 1 (perfect recall) |
| Specificity (98) | $\dfrac{TN}{FP + TN}$ | 0 (no specificity) – 1 (perfect specificity) |
| Quantity Disagreement (96) | $\dfrac{\left\lvert \dfrac{FN - FP}{TP + FP + FN + TN}\right\rvert + \left\lvert \dfrac{FP - FN}{TP + FP + FN + TN}\right\rvert}{2}$ | 0 (no disagreement) – 1 (complete disagreement) |
| Allocation Disagreement (96) | $2 * \min\left(\dfrac{FP}{TP + FP + FN + TN}, \dfrac{FN}{TP + FP + FN + TN}\right)$ | 0 (no disagreement) – 1 (complete disagreement) |
| F$_1$ score | $\dfrac{2 * \dfrac{TP}{TP + FP} * \dfrac{TP}{TP + FN}}{\dfrac{TP}{TP + FP} + \dfrac{TP}{TP + FN}}$ | 0 (worst) – 1 (best) |

As suggested by Pontius, Shusas, and McEachern (99), to assess whether the model is

adding any additional predictive ability , we constructed a naive (basic) model that randomly

assigns the transitions to a year within the given period, with every year having an equal

likelihood, and carried out predictions for each year within pixels that were known to have

transitioned for comparability with our model. Again, we determined for each pixel whether

it was a TP, FP, FN, or TN and calculated metrics to compare the BSGM and the naive model

at the across each country at the pixel level, and at the unit level. The naive model was

bootstrapped 500 times based upon resource limits and prediction stability, for each year and

was specific to each country.

## 3. RESULTS

Across all study areas, two-thirds of the modelled years correctly predicted between

85-99 percent of transition pixels. For all years, again at the pixel level, the BSGM displayed

low quantity and allocation disagreement in both absolute and relative terms. Similarly, the

pixel level F1 score, with few exceptions, was higher than the naive model, yet had more

variance in absolute terms of performance. Comparable results were found at the unit level,

with particularly good results in the middle and later years of the study period.

### 3.1 RF Performance

The ROC plots in Figure 2 show that the RFs approach the performance of the

theoretical perfect model. However, given the imbalanced data, the PRC plots (Figure 2)

show a more nuanced picture of performance where a maximum level of precision is quickly

achieved, remains steady up to a certain value of recall that varies by study area, and then

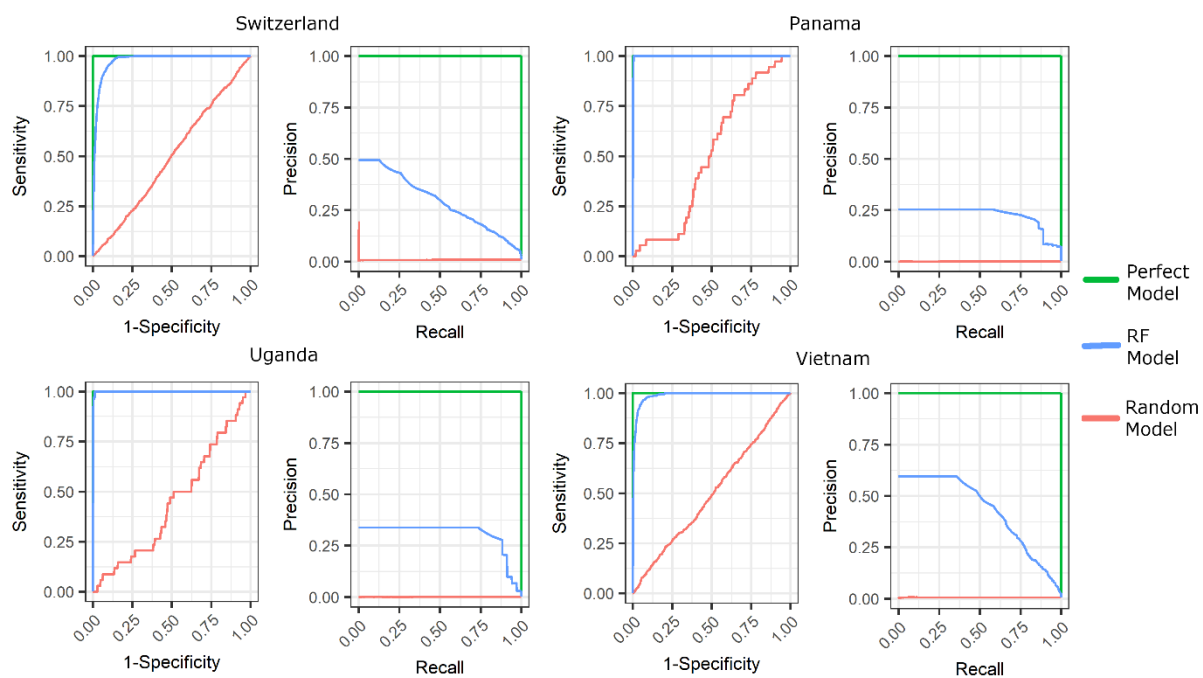precision quickly decreases with increasing recall.



**Figure 2.** Receiver Operator Curve (left plots) and Precision Recall Curves (right plots) with the RF model performance, blue lines, against a random model, red lines, and a perfect model, green lines, for each modelled country and input dataset.

Of the covariates informing the RFs of the models, we consistently saw that the most

important predictors of a pixel transitioning from non-BS to BS were covariates related to

distance ("esa_cls190_dst_2000") and local density ("esa_cls190_prp_5_2000",

"esa_cls190_prp_10_2000", and "esa_cls190_prp_15_2000") of BS established at the

21

beginning of the overall study period, i.e. 2000, and connectivity of BS extents at the beginning ("tt50k_2000") or end ("urbanaccessibility_2015" and "osmroa_dst") of the study period (Figure 3).
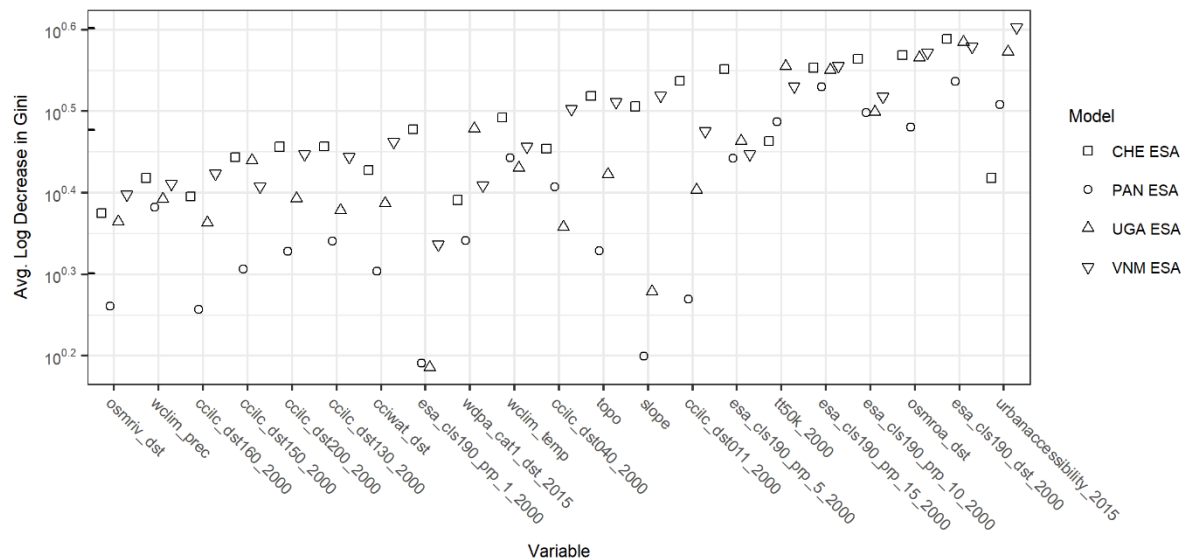


**Figure 3.** Random forest covariate importance as measured by the average log decrease in the Gini impurity when the covariate is used as the splitting criteria at nodes; higher values indicate better performance of covariate. Model for Swizerland (CHE) ESA, Panama (PAN) ESA, Uganda (UGA) ESA, and Vietnam (VNM) ESA, are shown. Refer to Table 2 for covariate names.

### 3.2 Pixel Level Results

Examining the proportion of pixels known to transition that were predicted correctly in Table 4, we show that out of 48 model years predicted, 39 of those years had correctly predicted proportions between 0.80 and 0.99 (green) with 25 of them having proportions over 0.90. The ESA based modelled years ranged from 0.57 to 0.99 of pixels predicted correctly (Table 4). Note that one minus the proportion correct is equal to the total disagreement of the predicted pixels, i.e. the sum of the quantity and allocation disagreement (96).

**Table 4.** Proportion of transition pixels predicted correctly by the BSGM by year. Note that 1 – the proportion correct is equal to the overall disagreement, i.e. the sum of the quantity and allocation disagreement. Modelled years with proportions greater than or equal to 0.80 are highlighted in green.

| Model | 2001 | 2002 | 2003 | 2004 | 2006 | 2007 | 2008 | 2009 | 2011 | 2012 | 2013 | 2014 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CHE ESA | 0.718 | 0.573 | 0.628 | 0.975 | 0.987 | 0.979 | 0.975 | 0.983 | 0.999 | 0.998 | 0.997 | 0.997 |
| PAN ESA | 0.952 | 0.935 | 0.934 | 0.960 | 0.806 | 0.771 | 0.816 | 0.920 | 0.905 | 0.838 | 0.801 | 0.818 |
| UGA ESA | 0.814 | 0.787 | 0.803 | 0.929 | 0.912 | 0.877 | 0.877 | 0.909 | 0.940 | 0.893 | 0.865 | 0.878 |

| VNM<br>ESA | 0.942 | 0.918 | 0.923 | 0.951 | 0.923 | 0.872 | 0.866 | 0.916 | 0.879 | 0.777 | 0.738 | 0.790 |

Examining the year-specific study area F1 scores in Figure 4, we show that the BSGM had low performance between 2001 and 2003, in absolute terms, and relatively near naive model performance, across all ESA study areas. The F1 score for all modelled years after 2003 increases to quite good performance, with values approaching 1.0 in some cases, and the difference between the BSGM and naive model performance increases even more dramatically (Figure 4).
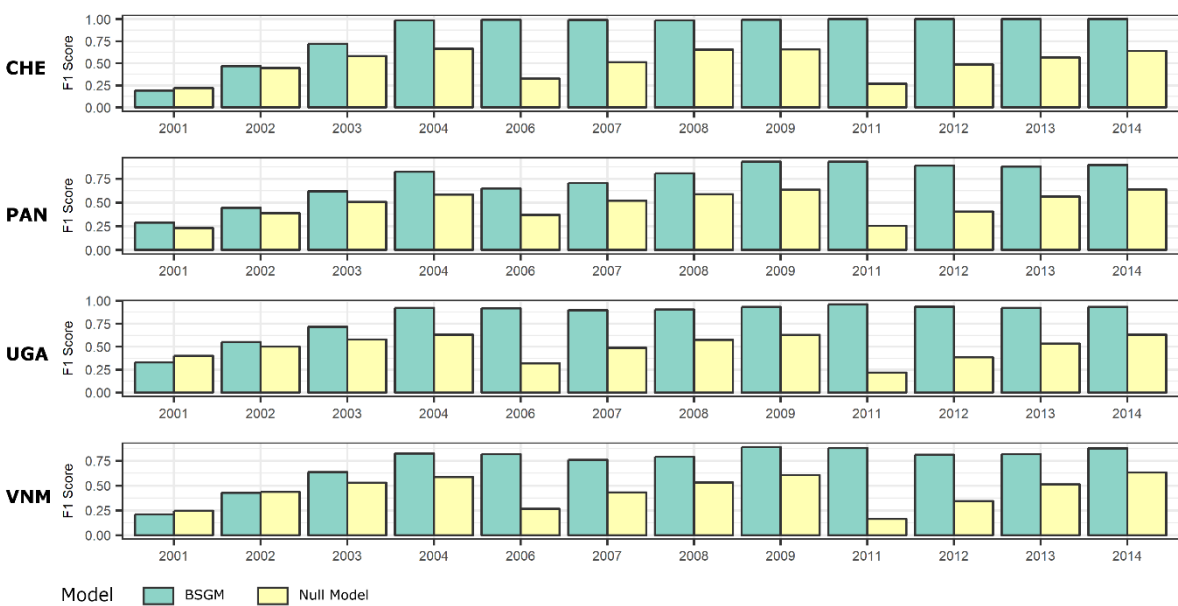


**Figure 4.** Pixel-level $F_1$ score by year for Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) as compared to a naive model. Full annual contingency data and metrics in supplementary material

Examining the source of the disagreement further, we display the observed quantity and allocation disagreement as well as the corresponding disagreements under the naive model in Figure 5. We show that for all modelled years using the ESA data, the total disagreement is substantially less than that of the naive model and predominantly, the disagreement produced by the BSGM model is predominantly due to allocation error (Figure 5). However, there does appear to be a pattern of increasing disagreement due to quantity error after 2010.
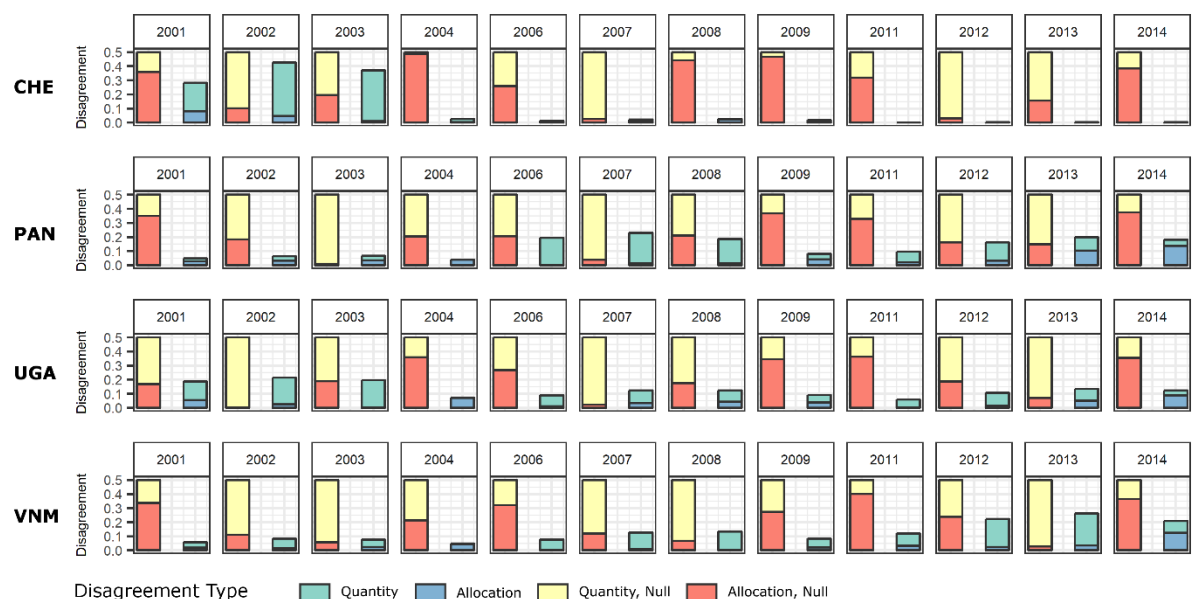
**Figure 5.** Pixel-level quantity and allocation disagreement of BSGM and naive models for Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) as compared to a naive model, given in red. Full annual contingency data and metrics in supplementary material

## 3.3 Subnational Unit Level Results

Overall, at the subnational unit level, we found results similar to the pixel-level results, including poor performance in absolute terms between 2001 to 2003, but some units were obviously performing worse than others as compared to the naive model. Plotting the ESA-informed model distributions of unit-level F1 scores by study area and year against the corresponding naive model performance, we show that the BSGM generally performs better in the majority of subnational units from which the transitions were disaggregated from (Figure 6). At worst, e.g. Vietnam 2002, approximately half of the units were still performing better than the naive model (Figure 6). For quantity disagreement (Figure 7) and allocation disagreement (Figure 8), results similar to pixel level results were found.
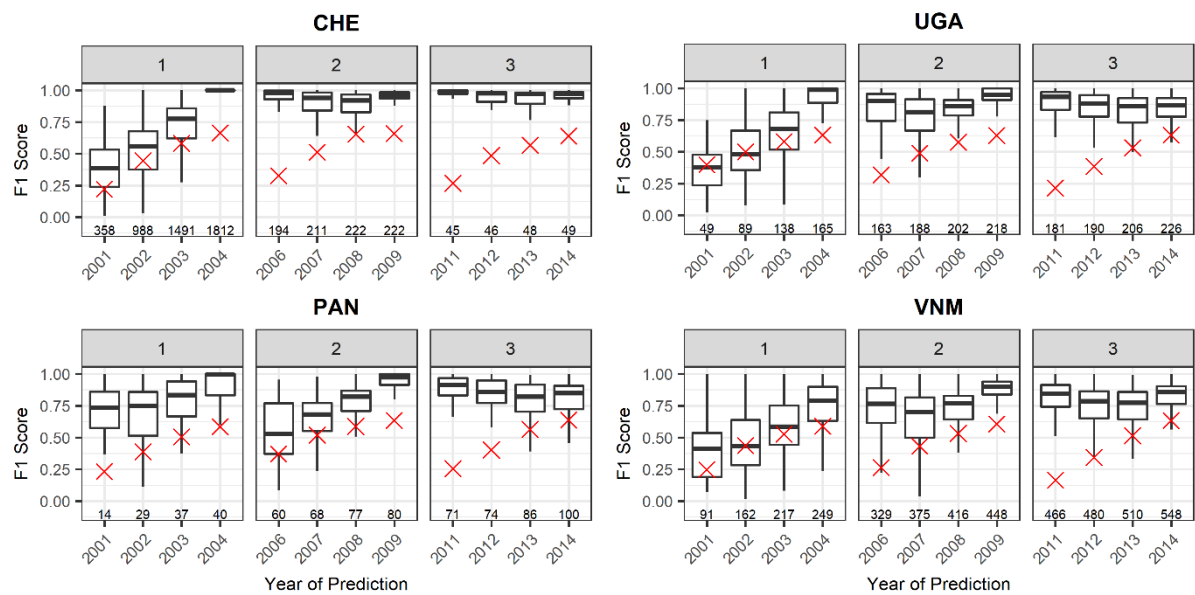
**Figure 6.** Unit level $F_1$ score box plots, by dasymetric period, of Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) ESA informed models as compared to a naive model, given by a red "x". Number of units exhibiting any transitions for each period and a defined metric value is given above the x-axis.
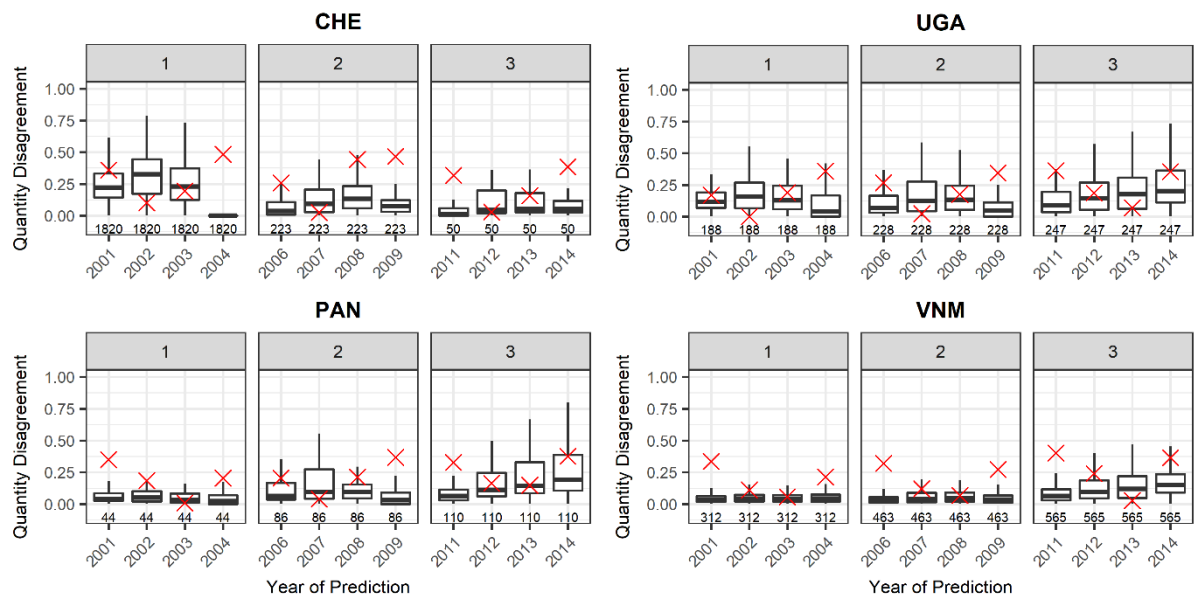


**Figure 7.** Unit level quantity disagreement box plots, by dasymetric period, of Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) ESA informed models as compared to a naive model, given by a red "x". Number of units exhibiting any transitions for each period and a defined metric value is given above the x-axis.

**Figure 8.** Unit level allocation disagreement box plots, by dasymetric period, of Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) ESA informed models as compared to a naive model, given by a red "x". Number of units exhibiting any transitions for each period and a defined metric value is given above the x-axis.
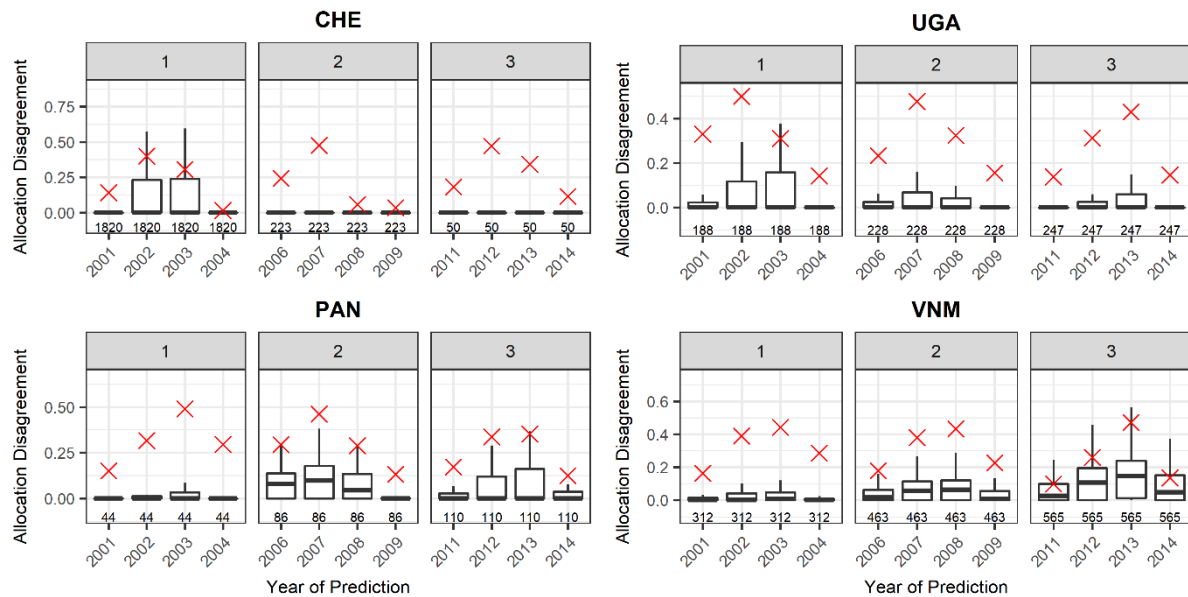
Plotting the unit-level metrics for all models as choropleth maps (see Supplementary Material for select maps and shape files containing contingency data), shows that years of generally good performance, the units of lesser performance are those that correspond to areas of less densely settled areas and the peripheries of established urban areas. Other years, such as Uganda 2001, performed poorly across many units with no apparent pattern. Identical analyses for the WSF Evolution data are given in the supplementary materials as well. Additionally, for each country the GeoTiffs showing the BS extents at 100m resolution for every predicted year are provided in the Supplementary Materials.

## 4. DISCUSSION

The 2030 Agenda for Sustainable Development and its SDGs, have reinforced the importance of data to being able to account for "all people everywhere (8). Differences in the dynamic spatial distributions of hazards (100,101), the spatial variation of the effects of climate change (5,102,103), spatially allocating services to ensure sufficient coverage

(104,105), and targeting interventions and planning (106,107) based upon local context with limited resources requires more accurate mapping of BS and mapping of populations, both large and small (8). Here we have shown a flexible modelling framework constructed from open-source methods and covariates to produce a framework that can be scaled to global extent across a variety of study areas and input data. This model approximates patterns of BS growth through time with good agreement for most years at the pixel and the units used in disaggregation (Table 4, Figures 4 and 6). Here we have shown the BSGM framework is capable of filling gaps of imagery-derived urban feature datasets by estimating the extents in between observations based upon relative changes in BS population and BS population density combined with environmental covariates. This emphasizes the strength of using an interpolative model, such as BSGM, as opposed to more imagery dependent annual feature extraction methods that may encounter adverse atmospheric conditions, limited sensor revisits, or the need for more resource intensive imagery-based interpolation or extraction methods. This framework, and resultant output data, can be used for better modelling of population through time, inform future urban feature extraction from imagery, help facilitate intervention/planning/monitoring of development goals, and potentially serve as a platform for simulating different transition paths through time and investigating correlates of BS spatial growth.

However, the BSGM is neither without error nor a total replacement for urban feature extractions. Given that the BSGM is interpolative, its predictions are limited by the accuracy, the spatial quality, and the temporal quality of its input urban feature dataset, the input time-specific subnational population data, and the input population surfaces. For example, the poorer model performance seen from 2001 through 2003 (Figures 4-8) are likely due to the fact the ESA data did not delineate changes, detected at 30 arc sec resolution, at 10 arc second resolution due to the MERIS and PROBA V imagery not being available (75).

With regards to total disagreement of the model (Figure 5), the relatively high contribution of allocation disagreement prior to circa 2010 and corresponding decrease in contribution post-2010 is likely due to the switch from using coarser DMSP-based LAN data to VIIRS-based LAN data at the 2012 time point.

Further, models are limited by their conceptual and mathematical assumptions. In this framework, we are assuming a certain relationship between relative population and population density changes and drive demand for temporally coincident settlement growth. This is not to say the assumed relationship is correct; here we assume BS population grows logistically with a time varying capacity that is temporally coincident, i.e. not lagged, and we assume BS population density follows a natural cubic spline across all observed points. This is further predicated upon the assumption that the BS growth is strongly correlated by changes in population and or population density and the resulting demand is instantaneously filled as opposed to being delayed temporally. While there is support for population change being an empirical and theoretical driver of urban/BS growth (20–22,42,58), there is also evidence for the use of other covariates, not used here because of their unavailability at subnational levels globally through time, such as Gross Domestic Product and arable land per capita (20,21). Furthermore, there are other "intangibles" such as local, regional, and national land use or development policies, which almost certainly shape BS growth, but are typically not in an accessible format, if available at all. The value of using population data to predict growth of settlement, shown here in a semi-independent model framework, and the value of using urban/settlement feature data sets to predict population distribution (108), raises the question of whether within a strictly predictive modelling framework (109,110), i.e. no inference on causality and measuring effect, if it is worthwhile or proper to try to fully separate population and settlement given their reciprocal causal links via economics, i.e.

economics drives population distribution and built-settlement which can then feed back into economics.

The modelling assumptions here are preceded by the assumption of exponential interpolation of annual population totals by the GPWv4-based data (66) and that the RF-informed dasymetric redistribution of those population totals are correctly locating the population in a manner which leads to correct BS population estimates for the BSGM to utilize. We already know that the RF population model tends to underestimate populations in BS and overestimate populations outside of urbanized areas (80). Since the BSGM allocates transitions based upon relative changes in BS population, this last point should not affect prediction timings, assuming the RF-informed population modelling, or any spatial population modelling approach, biases are consistent between the two times. Alternatively, any spatially explicit population datasets can be used as inputs for the BSGM, even the base GPWv4, removing the need to use a modelled population input. As with any "model built upon models," producers and users of such data must be cautious of accumulated data error, however, given the results here, error is relatively low for most predicted years, although some undoubtedly originates due to accumulated errors.

With any area-based metric the Modifiable Areal Unit Problem (97) must be considered, as the total number of pixels in each unit is typically larger in the less settled areas resulting in less variation of aggregated metric values in those areas. With dasymetric redistribution methods, the size and configuration of the source units, spatially or temporally, can also affect the quality of the disaggregation with the larger relative differences between source unit and target unit sizes introducing more uncertainty to the output (81,82). As the concept of urban growth, regardless of the chosen representation of urban, can be thought of as an incremental process, with future outcome dependent upon previous growth, the gridded

outputs of the BSGM can be aggregated across years to decrease uncertainty of the interpolated extents should annual datasets not be needed.

Unfortunately, for both the use of logistic curves to interpolate between estimates of built-settlement population data and cubic splines to interpolate between estimates of built-settlement population density, independent data does not exist to evaluate the error or uncertainty of these interpolated values largely because of absolute data availability. Further, even if such urban population and urban population density data existed, because of the differing definitions of urban (2,111,112), most probably we would have had a definitional disagreement with our adopted BS definition (unless these data came with corresponding spatial extents).

This aside, we also cannot calculate uncertainty of these curves because they are non-parametric growth curves or simple fitted splines that are not conducting any statistical inferences. However, we are not actually using the interpolated values of BS population and BS population density as the predicted outcome of interest, but rather to derive estimated counts of non-BS to BS pixel transitions that are then used as relative weights for the spatial disaggregation of the actual observed transition counts across time. This dasymetric disaggregation by weights precludes the propagation of any uncertainties calculated before the disaggregation step, a well-known characteristic of dasymetric methods, limiting us to only measuring absolute error of the final transitions as we have done here.

## 5. CONCLUSIONS

Here we introduced and validated a flexible modelling framework for globally modelling BS between observed time points, with 39 of 48 validated years having over 80 percent agreement and 25 of those years having over 90 percent agreement (Table 4). As global urban feature dataset producers such as ESA, MAUPP, GHSL, GUF and others

continue to improve and release datasets with higher temporal resolution, models such as the BSGM will likely still have utility due to imagery/extraction issues and a need to smooth or fill-in time periods where difficulty was experienced (39,40,62,113). By the time annual urban feature extractions from currently available imagery have become an economical means of filling gaps and the current demand for annual global urban extent datasets, eventually becoming the standard, there will have grown a demand for quarterly extents, monthly extents, and so on. Until such a time, globally consistent feature extractions and an interpolative model can attempt to fill the need, data permitting. This is not to say that interpolative models and imagery-extracted features are oppositional, but rather are complementary. Should the time come where high-resolution global annual feature datasets become the norm, this would offer a wealth of information from which to improve the model assumptions the BSGM currently makes. Further, the predictions of the BSGM could serve as a comparative check in the production of future urban feature extractions and a platform to explore population and BS dynamics.

As informative as global urban feature extraction datasets are, imagery will never see into the future and we plan on extending this framework to allow for short-term projection of BS growth, both in a predictive manner as well as allowing scenarios to be input. We found that the primary predictors of growth BS extents were related to connectivity, i.e. road networks, and local, i.e. ~0.5-1.5km, settlement density (Figure 3) both giving support to work in attempting to define "urban" base on contiguity, connectivity, and spatial density (114–116) and implying that investment in detecting/simulating new road network data would be beneficial to better predicting urban feature extents. Still mostly unknown is how the BSGM would perform for smaller settlements, not captured by coarser datasets such as the ESA land cover, and we are looking to test this with forthcoming feature data sets with resolutions below 3 arc seconds. Lastly, we plan to validate the utility of these dataset in an

applied manner by comparing the effects of including the BSGM derived extents in annual population modelling.

Here we have presented an open source framework for interpolating a binary BS or urban feature dataset using a limited covariate dataset that can be used to further population mapping by filling a current gap in annual global urban feature datasets. This framework is scalable globally, but also allows for sub-study area variation in transition probability, population changes, and lights-at-night changes to drive the overall study area model. Further, the model can be adapted to run at other scales, both spatially and temporally, either by modifying the provided code (See Supplementary Material) or in many cases, simply by modifying the input data. Annual global interpolated datasets from 2000 to 2014 based on GHSL/ESA/GUF input datasets, produced with an early version of this model and a reduced set of covariates, is freely available on the WorldPop website (worldpop.org) with the up to date model code used here provided in the supplementary material.

**REFERENCES**
1. Angel S, Sheppard SC, Civco DL. The Dynamics of Global Urban Expansion. Washington, D. C.: The World Bank; 2005.
2. United Nations. World Urbanization Prospects: The 2014 Revision [Internet]. New York, New York: United Nations, Dept. of Economic and Social Affairs, Population Division; 2015. (World Urbanization Prospects). Report No.: ST/ESA/SER.A/366. Available from: https://population.un.org/wup/Publications/Files/WUP2014-Methodology.pdf
3. Cohen B. Urbanization in Developing Countries: Current Trends, Future Projections, and Key Challenges for Sustainability. Technol Soc. 2006;28:63–80.
4. McGranahan G, Balk D, Anderson B. The Rising Tide: Assessing the Risks of Climate Change and Human Settlements in Low Elevation Coastal Zones. Environ Urban [Internet]. 2007;19(1):17–37. Available from: http://journals.sagepub.com/doi/pdf/10.1177/0956247807076960
5. Stephenson J, Newman K, Mayhew S. Population dynamics and climate change: What are the links? J Public Health (Bangkok). 2010;32(2):150–6.

6.    Dhingra MS, Artois J, Robinson TP, Linard C, Chaiban C, Xenarios I, et al. Global mapping of highly pathogenic avian influenza H5N1 and H5Nx clade 2.3.4.4 viruses with spatial cross-validation. Elife [Internet]. 2016 Nov 25;5. Available from: https://elifesciences.org/articles/19571

7.    Chongsuvivatwong V, Phua KH, Yap MT, Pocock NS, Hashim JH, Chhem R, et al. Health and health-care systems in southeast Asia: diversity and transitions. Lancet [Internet]. 2011 Jan;377(9763):429–37. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0140673610615073

8.    United Nations. Transforming Our World: The 2030 Agenda for Sustainable Development [Internet]. 2016. Available from: https://sustainabledevelopment.un.org/content/documents/21252030 Agenda for Sustainable Development web.pdf

9.    UN. World Population Prospects: The 2014 Revision. Washington, D. C.; 2015.

10.   United Nations. World Urbanization Prospects: The 2018 Revision [Internet]. New York; 2018. Available from: https://population.un.org/wup/Publications/Files/WUP2018-Methodology.pdf

11.   Potere D, Schneider A. A critical look at representations of urban areas in global maps. GeoJournal [Internet]. 2007 Oct 10;69(1–2):55–80. Available from: http://link.springer.com/10.1007/s10708-007-9102-z

12.   Merriam-Webster. urban [Internet]. Merriam-Webster Dictionary. 2019 [cited 2019 Feb 7]. Available from: https://www.merriam-webster.com/dictionary/urban

13.   Burgess EW. The Growth of a City: An Introduction to a Research Project. Chicago, IL: University of Chicago Press; 1925.

14.   Southworth F. ORNL-6881: A Technical Review of URban Land Use-Transportation Models as a Tool for Evaluating Vehicle Travel Reduction Strategies. Oak Ridge, TN; 1995.

15.   Parr J. The polycentric urban region: A closer inspection. Reg Stud. 2004;38(3):231–40.

16.   Cohen B. Urban growth in developing countries: A review of current trends and a caution regarding existing forecasting. World Dev. 2004;32(1):23–51.

17.   Pozzi F, Small C. Analysis of urban land cover and population density in the United States. Photogramm Eng Remote Sensing2. 2005;71:719–26.

18.   Schneider A, Friedl MA, Potere D. Mapping Urban Areas Using MODIS 500-m Data: New Methods and Datasets Based on "Urban Ecoregions." Remote Sens Environ. 2010;114:1733–46.

19.   Haas H De. Migration and Development: A Theoretical Perspective. Int Migr Rev. 2010;44(1):227–64.

20.   Angel S, Parent J, Civco DL, Blei AM, Potere D. The Dimensions of Global Urban Expansion: Estimates and Projections for All Countries, 2000-2050. Prog Plann. 2011;75:53–107.

21.   Seto KC, Fragkias M, Guneralp B, Reilly MK. A Meta-Analysis of Global Urban Land Expansion. PLoS One. 2011;6(8):e23777.

22.   Dyson T. The role of the demographic transition in the process of urbanization. Popul Dev Rev. 2011;37(Supplement):34–54.

23.   Hoyt H. The Structure and Growth of Residential Neighborhoods in American Cities. Washington, D. C.: United States Government Printing Office; 1939.

24.   Harris CD, Ullman EL. The nature of cities. Ann Am Acad Polit Soc Sci. 1945;242:7–17.

25.   Gottman J. Megalopolis, or the urbanisation of the north eastern seaboard. Econ Geogr. 1957;33:189–200.

26. Von Thunen JH. Von Thunen's "Isolated State": An English Translation of "Der Isolierte Staat." Wartenberg CM, Hall P, editors. Oxford, UK: Pergamon Press; 1966.

27. Zelinsky W. The Hypothesis of the Mobility Transition. Geogr Rev. 1971;61(2):219–49.

28. Ledent J. Rural-Urban Migration, Urbanization, and Economic Development. Econ Dev Cult Change [Internet]. 1982;30(3):507–38. Available from: https://www.jstor.org/stable/3203205

29. Berechman J, Gordon P. Linked Models of Land-Use Transport Interactions: A Review. In: Hutchinson B, Batty M, editors. Advances in Urban Systems Modelling. Elsevier Ltd; 1986.

30. Meyer WB, Turner BL. Human Population Growth and Global Land-Use / Cover Change. Annu Rev Ecol Syst. 1992;23(1992):39–61.

31. Schneider A, Woodcock CE. Compact, Dispersed, Fragmented, Extensive? A Comparison of Urban Growth in Twenty-five Global Cities using Remotely Sensed Data, Pattern Metrics and Census Information. Urban Stud [Internet]. 2008 Mar;45(3):659–92. Available from: http://journals.sagepub.com/doi/10.1177/0042098007087340

32. Small C. The color of cities:An overview of urban spectral diversity. In: Herold M, Gamba P, editors. Global Mapping of Human Settlements. New York: Taylor & Francis; 2009. p. 59–106.

33. Bartholomé E, Belward AS. GLC2000: a new approach to global land cover mapping from Earth observation data. Int J Remote Sens [Internet]. 2005 May;26(9):1959–77. Available from: http://www.tandfonline.com/doi/abs/10.1080/01431160412331291297

34. Schneider A, Friedl MA, McIver DK, Woodcock CE. Mapping Urban Areas by Fusing Multiple Sources of Coarse Resolution Remotely Sensed Data. Photogramm Remote Sens. 2003;69(12):1377–86.

35. Potere D, Schneider A, Angel S, Civco D. Mapping urban areas on a global scale: which of the eight maps now available is more accurate? Int J Remote Sens [Internet]. 2009;30(24):6531–58. Available from: http://www.tandfonline.com/doi/abs/10.1080/01431160903121134

36. Huang X, Wen D, Li J, Qin R. Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery. Remote Sens Environ [Internet]. 2017;196:56–75. Available from: http://dx.doi.org/10.1016/j.rse.2017.05.001

37. Liu X, Hu G, Chen Y, Li X, Xu X, Li S, et al. High-resolution multi-temporal mapping of global urban land using Landsat images based on the Google Earth Engine Platform. Remote Sens Environ [Internet]. 2018;209:227–39. Available from: https://doi.org/10.1016/j.rse.2018.02.055

38. Pesaresi M, Guo H, Blaes X, Ehrlich D, Ferri S, Gueguen L, et al. A Global Human Settlement Layer from Optical HR/VHR Remote Sensing Data: Concept and First Results. IEEE J Sel Top Appl Earth Obs Remote Sens. 2013;6(5):2102–31.

39. Pesaresi M, Ehrlich D, Ferri S, Florczyk AJ, Freire S, Halkia S, et al. Operating Procedure for the Production of the Global Human Settlement Layer from Landsat Data of the Epochs 1975, 1990, 2000, and 2014 [Internet]. Publications Office of the European Union; 2016. Available from: doi: 10.2788/253582

40. Esch T, Marconcini M, Felbier A, Roth A, Heldens W, Huber M, et al. Urban Footprint Processor - Fully Automated Processing Chain Generating Settlement Masks from Global Data of the TanDEM-X Mission. IEEE Geosci Remote Sens Lett. 2013;10(6):1617–21.

41. Esch T, Bachofer F, Heldens W, Hirner A, Marconcini M, Palacios-Lopez D, et al.

Where We Live—A Summary of the Achievements and Planned Evolution of the Global Urban Footprint. Remote Sens [Internet]. 2018 Jun 7;10(6):895. Available from: http://www.mdpi.com/2072-4292/10/6/895

42.     Linard C, Tatem AJ, Gilbert M. Modelling Spatial Patterns of Urban Growth in Africa. Appl Geogr. 2013;44:23–32.

43.     Barredo JI, Demicheli L, Lavalle C, Kasanko M, McCormick N. Modelling Future Urban Scenarios in Developing Countries: An Application Case Study in Lagos, Nigeria. Environ Plan B. 2004;31:65–84.

44.     Batty M, Xie Y. From Cells to Cities. Environ Plan B. 1994;21:S31–48.

45.     Clarke KC, Hoppen S, Gaydos L. A Self-modifying Cellular Automaton Model of Historical Urbanization in the San Francisco Bay Area. Environ Plan B. 1997;24:247–61.

46.     Clarke KC, Gaydos L. Loose-coupling a Cellular Automaton Model and GIS: Long-term Urban Growth Prediction for San Francisco and Washington/Baltimore. Int J Geogr Inf Sci. 1998;12(7):699–714.

47.     Leao S, Bishop I, Evans D. Simulating Urban Growth in a Developing Nation's Region Using a Cellular Automata-based Model. J Urban Plan Dev. 2004;130(3):145–58.

48.     White R, Engelen G. High Resolution Modelling of the Spatial Dynamics of Urban and Regional Systems. Comput Environ Urban Syst. 2000;24(383–400).

49.     White R, Engelen G. Cellular Automata as the Basis of Integrated Dynamic Regional Modelling. Environ Plan B. 1997;24:235–46.

50.     Sante I, Garcia AM, Miranda D, Crecente R. Cellular Automata Models for the Simulation of Real-world Urban Processes: A Review and Analysis. Landsc Urban Plan. 2010;96:108–22.

51.     Huang X, Hu T, Li J, Wang Q. Mapping Urban Areas in China Using Multisource Data With a Novel Ensemble SVM Method. IEEE Trans Geosci Remote Sens. 2018;56(8):4258–73.

52.     Verburg PH, Schot PP, Dijst MJ, Veldkamp A. Landuse Change Modelling: Current Practice and Research Priorities. GeoJournal. 2004;61:309–24.

53.     Tayyebi A, Pekin BK, Pijanowski BC, Plourde JD, Doucette JS, Braun D. Hierarchical modeling of urban growth across the conterminous USA: Developing meso-scale quantity drivers for the Land Transformation Model. J Land Use Sci. 2013;8(4):422–42.

54.     Batty M. Urban Modeling. In: International Encyclopedia of Human Geography. Oxford, UK: Elsevier; 2009. p. 51–8.

55.     Verburg PH, Soepboer W, Veldkamp A, Limpiada R, Espladon V, Mastura SSA. Modeling the Spatial Dynamics of Regional Land Use: The CLUE-S Model. Environ Manage [Internet]. 2002 Sep 9;30(3):391–405. Available from: http://link.springer.com/10.1007/s00267-002-2630-x

56.     Patel N, Angiuli E, Gamba P, Gaughan AE, Lisini G, Stevens FR, et al. Multitemporal Settlement and Population Mapping From Landsat Using Google Earth Engine. Int J Appl Earth Obs Geoinf [Internet]. 2015;35(Part B):199–208. Available from: http://dx.doi.org/10.1016/j.jag.2014.09.005

57.     Li X, Gong P. Urban growth models: progress and perspective. Sci Bull [Internet]. 2016 Nov;61(21):1637–50. Available from: https://linkinghub.elsevier.com/retrieve/pii/S2095927316305540

58.     Seto KC, Guneralp B, Hutyra LR. Global Forecasts of Urban Expansion to 2030 and Direct Impacts on Biodiversity and Carbon Pools. Proc Natl Acad Sci U S A. 2012;109(40):16083–8.

59.     Goldewijk KK, Beusen A, Janssen P. Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1. The Holocene. 2010;20(4):565–73.

60.     Tobler W, Deichmann U, Gottsegen J, Maloy K. World Population in a Grid of Spherical Quadrilaterals. Int J Popul Geogr. 1997;3:203–25.

61.     Verburg PH, de Koning GHJ, Kok K, Veldkamp A, Bouma J. A spatial explicit allocation procedure for modelling the pattern of land use change based upon actual land use. Ecol Modell. 1999;116:45–61.

62.     ESA CCI. European Space Agency Climate Change Initiative Landcover [Internet]. v 2.0.7. European Space Agency; 2017. Available from: http://maps.elie.ucl.ac.be/CCI/viewer/download.php

63.     Lehner B, Verdin K, Jarvis A. New Global Hydrography Derived from Spaceborne Elevation Data. Eos, Trans Am Geophys Union. 2008;89(10):93–4.

64.     U.N. Enviroment Programme World Conservation Monitoring Centre, IUCN World Commission on Protected Areas. World Database on Protected Areas [Internet]. 2nd ed. IUCN & UNEP; 2015. Available from: https://www.protectedplanet.net/

65.     Lamarche C, Santoro M, Bontemps S, D'Andrimont R, Radoux J, Giustarini L, et al. Compilation and Validation of SAR and Optical Data Products for a Complete and Global Map of Inland/Ocean Water Tailored to the Climate Modeling Community. Remote Sens [Internet]. 2017 Jan 11;9(36). Available from: http://www.mdpi.com/2072-4292/9/1/36

66.     Doxsey-Whitfield E, MacManus K, Adamo SB, Pistolesi L, Squires J, Borkovska O, et al. Taking advantage of the improved availability of census data: A first look at the Gridded Population of the World, Version 4. Pap Appl Geogr. 2015;1(3):226–34.

67.     Zhang Q, Pandey B, Seto KC. A Robust Method to Generate a Consistent Time Series From DMSP/OLS Nighttime Light Data. IEEE Trans Geosci Remote Sens [Internet]. 2016 Oct;54(10):5821–31. Available from: http://ieeexplore.ieee.org/document/7490418/

68.     WorldPop S of G and ESU of, Department of Geography and Geosciences U of L, Département de Géographie U de N, Center for International Earth Science Information Network (CIESIN) CU. Global High Resolution Population Denominators Project. Bill and Melinda Gates Foundation (OPP1134076); 2018.

69.     Earth Observation Group NOAA National Geophysical Data Center. VIIRS Nighttime Lights - One Month Composites [Internet]. Boulder, Colorado: NOAA National Centers for Environmental Information; 2016. Available from: https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html

70.     Nelson A. Estimated Travel Time to the Nearest city of 50,000 or More People in Year 2000 [Internet]. Ispra, Italy: Global Environment Monitoring Unit - Joint Research Centre of the European Commission; 2008. Available from: http://forobs.jrc.ec.europa.eu/products/gam/sources.php

71.     Weiss DJ, Nelson A, Gibson HS, Temperley W, Peedell S, Lieber A, et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature [Internet]. 2018;553(7688):333–6. Available from: http://dx.doi.org/10.1038/nature25181

72.     OpenStreetMap Contributers. OpenStreetMap (OSM) Database [Internet]. OSM; 2017. Available from: openstreetmap.org

73.     Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. Very high resolution interpolated climate surfaces for global land areas. Int J Climatol. 2005;25:1965–78.

74.     Sorichetta A, Hornby GM, Stevens FR, Gaughan AE, Linard C, Tatem AJ. High-resolution gridded population distribution datasets of Latin America in 2010, 2015,

and 2020. Sci Data. 2015;2:150045.

75. UCL Geomatics. Land Cover CCI Product User Guide Version 2.0 [Internet]. 2017. Available from: http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf

76. Breiman L. Random Forests. Mach Learn. 2001;45(1):5–32.

77. R Core Team. R: A Language and Environment Layer for Statistical Computing [Internet]. 3.3.1. Vienna, Austria: R Foundation for Statistical Computing; 2016. Available from: https://www.r-project.org

78. He H, Garcia EA. Learning from Imbalanced Data. 2009;21(9):1263–84.

79. Kamusoko C, Gamba J. Simulating Urban Growth Using a Random Forest-Cellular Automata (RF-CA) Model. ISPRS Int J Geo-Information. 2015;4:447–70.

80. Stevens FR, Gaughan AE, Linard C, Tatem AJ. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-sensed Data and Ancillary Data. PLoS One. 2015;10(2):e0107042.

81. Mennis J. Generating surface models of population using dasymetric mapping. Prof Geogr. 2003;55(1):31–42.

82. Mennis J, Hultgren T. Intelligent dasymetric mapping and its application to areal interpolation. Cartogr Geogr Inf Sci. 2006;33:179–94.

83. Gaughan AE, Stevens FR, Huang Z, Nieves JJ, Sorichetta A, Lai S, et al. Spatiotemporal Patterns of Population in Mainland China, 1990 to 2010. Sci Data. 2016;3:160005.

84. Booth H. Demographic forecasting: 1980 to 2005 in review. Int J Forecast [Internet]. 2006 Jan;22(3):547–81. Available from: http://linkinghub.elsevier.com/retrieve/pii/S016920700600046X

85. Meyer PS, Ausubel JH. Carrying capacity: A model with logistically varying limits. Technol Forecast Soc Change. 1999;61(3):209–14.

86. Austin AL, Brewer JW. World Population Growth and Related Technical Problems. Technol Forecast Soc Chang. 1971;3(1):23–49.

87. Cohen JE. Population Growth and Earth's Human Carrying Capacity. Science (80- ). 1995;269(5222):341–6.

88. Wilson AG. Catastrophe Theory and Urban Modelling: An Application to Modal Choice. Environ Plan A. 1976;8(3):351–6.

89. Smith SK. Further thoughts on simplicity and complexity in population projection models. Int J Forecast. 1997;13:557–65.

90. Sibly RM. On the Regulation of Populations of Mammals, Birds, Fish, and Insects. Science (80- ) [Internet]. 2005 Jul 22;309(5734):607–10. Available from: http://www.sciencemag.org/cgi/doi/10.1126/science.1127705

91. Verhulst P-F. Notice sur la loi que la population poursuit dans son accroissement. Corresp Mathe-matique Phys. 1838;10:113–21.

92. McNeil DR, Trussell TJ, Turner JC. Spline Interpolation of Demographic Data. Demography [Internet]. 1977;14(2):245–52. Available from: https://www.jstor.org/stable/2060581

93. Saito T, Rehmsmeier M. The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. PLoS One. 2015;10(3):e0118432.

94. Haibo He, Garcia EA. Learning from Imbalanced Data. IEEE Trans Knowl Data Eng [Internet]. 2009 Sep;21(9):1263–84. Available from: http://ieeexplore.ieee.org/document/5128907/

95. Pontius RG, Millones M. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int J Remote Sens [Internet]. 2011

Aug 10;32(15):4407–29. Available from:
https://www.tandfonline.com/doi/full/10.1080/01431161.2011.552923

96.  Pontius RG, Millones M. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. Int J Remote Sens. 2011 Aug;32(15):4407–29.

97.  Openshaw S. The modifiable areal unit problem. Concepts Tech Mod Geogr. 1984;38.

98.  Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. Am J Epidemiol. 1978;107(1):71–6.

99.  Pontius RG, Shusas E, McEachern M. Detecting important categorical land changes while accounting for persistence. Agric Ecosyst Environ. 2004 Feb;101(2–3):251–68.

100. Carrão H, Naumann G, Barbosa P. Mapping global patterns of drought risk: An empirical framework based on sub-national estimates of hazard, exposure and vulnerability. Glob Environ Chang. 2016;39:108–24.

101. Oliveira S, Oehler F, San-Miguel-Ayanz J, Camia A, Pereira JMC. Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. For Ecol Manage [Internet]. 2012;275(July 2012):117–29. Available from: http://dx.doi.org/10.1016/j.foreco.2012.03.003

102. Ericson JP, Vorosmarty CJ, Dingman SL, Ward LG, Meybeck M. Effective sea-level rise and deltas: Causes of change and human dimension implications. Glob Planet Change. 2006;50:63–82.

103. Hanjra MA, Qureshi ME. Global Water Crisis and Future Food Security in an Era of Climate Change. Food Policy. 2010;35:365–77.

104. Sverdlik A. Ill-health and poverty: A literature review on health in informal settlements. Environ Urban. 2011;23(1):123–55.

105. Eckert S, Kohler S. Urbanization and Health in Developing Countries: A Systematic Review. World Health Popul. 2014;15(1):7–20.

106. Utazi CE, Thorley J, Alegana VA, Ferrari MJ, Takahashi S, Metcalf CJE, et al. High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. Vaccine [Internet]. 2018 Mar;36(12):1583–91. Available from: http://linkinghub.elsevier.com/retrieve/pii/S0264410X18301944

107. Linard C, Alegana V, Noor AM, Snow RW, Tatem AJ. A high resolution spatial population database of somolia for disease risk mapping. Int J Health Geogr. 2010;9(1):45.

108. Nieves JJ, Stevens FR, Gaughan AE, Linard C, Sorichetta A, Hornby G, et al. Examining the correlates and drivers of human population distributions across low- and middle-income countries. J R Soc Interface [Internet]. 2017 Dec 13;14(137):20170401. Available from: http://rsif.royalsocietypublishing.org/lookup/doi/10.1098/rsif.2017.0401

109. Breiman L. Statistical Modeling: The Two Cultures. Stat Sci. 2001;16(3):199–231.

110. Shmueli G. To Explain or Predict. Stat Sci. 2010;25(3):289–310.

111. Balk D, Leyk S, Jones B, Montgomery MR, Clark A. Understanding urbanization: A study of census and satellite-derived urban classes in the United States, 1990-2010. Benenson I, editor. PLoS One. 2018 Dec;13(12):e0208487.

112. Balk D, Montgomery M, Engin H, Lin N, Major E, Jones B. Urbanization in India: Population and Urban Classification Grids for 2011. Data. 2019 Feb;4(1):35.

113. Forget Y, Linard C, Gilbert M. Supervised Classification of Built-Up Areas in Sub-Saharan African Cities Using Landsat Imagery and OpenStreetMap. Remote Sens [Internet]. 2018 Jul 20;10(7):1145. Available from: http://www.mdpi.com/2072-4292/10/7/1145

114. Esch T, Marconcini M, Marmanis D, Zeidler J, Elsayed S, Metz A, et al.

Dimensioning urbanization - An advanced procedure for characterizing human settlement properties using spatial network analysis. Appl Geogr. 2014;55:212–28.

115. Pesaresi M, Freire S. GHS settlement grid, following the REGIO model 2014 in application to GHSL Landsat and CIESIN GPW v4-multitemporal (1975-1990-2000-2015) [Internet]. European Commission, Joint Research Centre [Dataset]. 2016 [cited 2018 Oct 26]. Available from: http://data.jrc.ec.europa.eu/dataset/jrc-ghsl-ghs_smod_pop_globe_r2016a

116. Dijkstra L, Poelman H. A harmonized definition of cities and rural areas: the new degree of urbanization [Internet]. 2014. (Regional Working Paper). Report No.: WP 01/2014. Available from:
http://ec.europa.eu/regional_policy/sources/docgener/work/2014_01_new_urban.pdf