INTERFACE

rsif.royalsocietypublishing.org

Research

Cite this article: Nieves JJ, Stevens FR, Gaughan AE, Linard C, Sorichetta A, Hornby G, Patel NN, Tatem AJ. 2017 Examining the correlates and drivers of human population distributions across low- and middle-income countries. *J. R. Soc. Interface* **14**: 20170401. http://dx.doi.org/10.1098/rsif.2017.0401

Received: 31 May 2017
Accepted: 20 November 2017

Subject Category:
Life Sciences–Earth Science interface

Subject Areas:
environmental science

Keywords:
population, mapping, census, dasymetric, disaggregation, random forests

Author for correspondence:
Jeremiah J. Nieves
e-mail: jeremiah.j.nieves@gmail.com

†Present address: Geography and Environment, University of Southampton, Building 44, Room 54/2001, University Road, Southampton SO17 1BJ, UK.

Electronic supplementary material is available online at https://dx.doi.org/10.6084/m9.figshare.c.3941356.

THE ROYAL SOCIETY PUBLISHING

Examining the correlates and drivers of human population distributions across low- and middle-income countries

Jeremiah J. Nieves[1,†], Forrest R. Stevens[1], Andrea E. Gaughan[1], Catherine Linard[2,3], Alessandro Sorichetta[4,5], Graeme Hornby[6], Nirav N. Patel[7] and Andrew J. Tatem[4,5]

[1]Department of Geography and Geosciences, University of Louisville, Lutz Hall, Louisville, KY 40292, USA
[2]Department of Geography, Université de Namur, Rue de Bruxelles 61, 5000 Namur, Belgium
[3]Spatial Epidemiology Lab (SpELL), Université Libre de Bruxelles CP160/12, Avenue F.D. Roosevelt 50, 1050 Brussels, Belgium
[4]WorldPop, Geography and Environment, University of Southampton, Building 44, Room 54/2001, University Road, Southampton SO17 1BJ, UK
[5]Flowminder Foundation, Stockholm, Sweden
[6]GeoData, University of Southampton, Building 44, Room 44/2087, University Road, Southampton SO17 1BJ, UK
[7]Department of Geography and Geoinformation Science, George Mason University, 4400 University Drive, MS 6C3, Fairfax, VA 22030, USA

JJN, 0000-0002-7423-1341

Geographical factors have influenced the distributions and densities of global human population distributions for centuries. Climatic regimes have made some regions more habitable than others, harsh topography has discouraged human settlement, and transport links have encouraged population growth. A better understanding of these types of relationships enables both improved mapping of population distributions today and modelling of future scenarios. However, few comprehensive studies of the relationships between population spatial distributions and the range of drivers and correlates that exist have been undertaken at all, much less at high spatial resolutions, and particularly across the low- and middle-income countries. Here, we quantify the relative importance of multiple types of drivers and covariates in explaining observed population densities across 32 low- and middle-income countries over four continents using machine-learning approaches. We find that, while relationships between population densities and geographical factors show some variation between regions, they are generally remarkably consistent, pointing to universal drivers of human population distribution. Here, we find that a set of geographical features relating to the built environment, ecology and topography consistently explain the majority of variability in population distributions at fine spatial scales across the low- and middle-income regions of the world.

1. Introduction

While archaeologists have long stated that settlement patterns are complex and multi-factorial, geography has always been a determinant of the location of human settlements with humans primarily settling where resources are available, such as coastal areas and arable lands [1–5]. Access to sufficient resources to meet the needs of a population limit the population densities in any given location while other locations may have climates and topography that are less conducive to supporting human populations. However, the location of human populations is not simply determined by the natural environment, i.e. environmental determinism [6]. Since the agricultural revolution, humans have often been the drivers of change in the natural environment, modifying it in ways to better access resources/services (e.g. transportation networks, densification of services

and production in urban areas) or to make the natural environment more productive and habitable (e.g. land conversion to agriculture, wetland drainage, irrigation, shelter in settlements) [7–11]. Sometimes humans have modified the environment in ways that make it less habitable, such as through pollution and desertification, or no longer habitable, such as in the cases of radiation in areas surrounding Chernobyl or desiccation of the Aral Sea [8,12,13]. With these changes, settlements and urban areas and populations continue to grow and their spatial distributions continue to evolve [14–16].

Between 2015 and 2050, the UN estimates that the global human population will grow by 2.4 billion [17]. Most of this projected change is anticipated to occur in the least developed countries and in urbanized areas [15,16]. Concurrently, Africa, Asia, Latin America and the Caribbean are estimated to experience the highest rates of urbanization [15]. As a part of this 'urban transition', the majority of Africa and Asia are experiencing large rates of internal migration, international migration and changes in the spatial distribution of natural population growth [15,16]. While Latin America and the Caribbean are predicted to experience decreasing urbanization rates, as was the trend through the 1990s and the early 2000s, the region is expected to have major demographic shifts. These rapidly changing magnitudes, composition and distribution of human populations imply a continued if not increasing need for high-resolution spatially explicit population maps that more accurately capture these changes to facilitate public health, sustainability and policy planning in general.

Over the past 20 years, the advancement of statistical techniques, availability of consistent geospatial data and rise in processing power have been leveraged to more accurately map populations over global scales. Such efforts include the simple gridding of census data matched to administrative boundaries that is undertaken for the Gridded Population of the World (GPW) project [18], and the use of satellite images of night-time lights to map urban areas and allocate populations to them, in the case of the Global Rural Urban Mapping Project (GRUMP) [19–21]. Other ongoing efforts, including LandScan [22–24], the Global Human Settlement Population Grid (GHS-POP) [25] and WorldPop [26], focus on a multivariate approach, utilizing multiple geospatial layers representing factors related to human population distributions to disaggregate areal unit-based census population counts to fine spatial resolution grid squares. These approaches can assess the contribution of different factors in explaining the observed population distributions (e.g. [26]), providing valuable data on the drivers and correlates of these patterns.

Despite the development of these multivariate approaches, there have been few globally representative comprehensive studies on the relationships between population densities, their associated covariates and the ancillary datasets that represent the covariates at a sub-national scale. Only basic within-country analyses have been undertaken in the course of validation or accuracy assessment, yet no analysis across low- and middle-income countries has occurred [26–29]. However, some local-scale case studies have investigated associations between covariates and population or residential land to better understand the correlates and drivers of population distributions in different settings [30,31]. Additionally, dasymetric modelling has evolved significantly over the past few years and provided important insights into the relationships between population and ancillary variables [32–35]. Such analyses have the potential to uncover fundamental patterns in the correlates and drivers of population distributions across the world.

Here, we undertake such an analysis for 32 low- and middle-income countries, focusing on answering the following two questions. (i) What datasets, representing drivers and associated landscapes of population distribution, are the most informative for accurately mapping populations at global scales?; (ii) What are the differences, in terms of relative importance of these datasets, between countries, between regions of countries and within regions of countries? By quantifying the relative importance of the drivers and correlates of human population distributions in relation to observed population densities, the question of how populations are distributed, and how this varies geographically, can begin to be addressed. Furthermore, it will allow informed development of new ancillary datasets with a high probability of importance when placed within a modelling framework and potentially lead to more informed covariate choices in population modelling that can expand the possible end-use applications of the population data. Moreover, by better depicting the relative importance of the drivers and associated landscapes of populations at the global and regional scales the accuracy and precision of high-resolution population mapping and construction of future scenarios will be furthered, benefitting all down-stream applications.

2. Material and methods

To assess the relationships between population densities and candidate correlates and drivers, we built a machine-learning-based modelling framework to expose the relationships between sub-national boundary-matched population census data and a library of geospatial datasets. The population models considered in this study are based on the random forest (RF)-based method as described in Stevens *et al*. [26]. We took the RF regression model objects for each sample country which were trained at the administrative unit level of the corresponding census-based population data, extracted the covariate importance metrics, standardized what the covariates were representing to facilitate comparisons across models and analysed these data for differences between and within covariate classes as well as within each covariate class between all countries, between regions and within regions to begin to address the possibility of geographic variability in these relationships.

2.1. Random forest-based population models

RFs are a non-parametric, nonlinear statistical method that falls within a category of machine-learning methods known as 'ensemble methods'. Ensemble methods take individual decision trees that are considered 'weak learners' and combine them to create a 'strong learner'. The benefits of ensemble methods are that generalizability is increased, performance on large or small datasets is improved and the ability of the method to model difficult learning tasks is more effective. Compared with other ensemble methods RFs are robust to noise, small sample sizes and over-fitting, yet they need little in the way of parameter specifications [36–39].

RFs independently generate k number of unpruned decision trees using 'bagging' [37,40]. Once a decision tree is grown, the one-third of the bagged training data that the tree was not grown upon remain and are known as the 'out-of-bag' (OOB) data. The decision tree applied to these data and the accuracy of the tree, as measured by the mean squared error (m.s.e.), are stored as the OOB error for that tree [37]. The prediction error of the entire RF model can be estimated by averaging the OOB

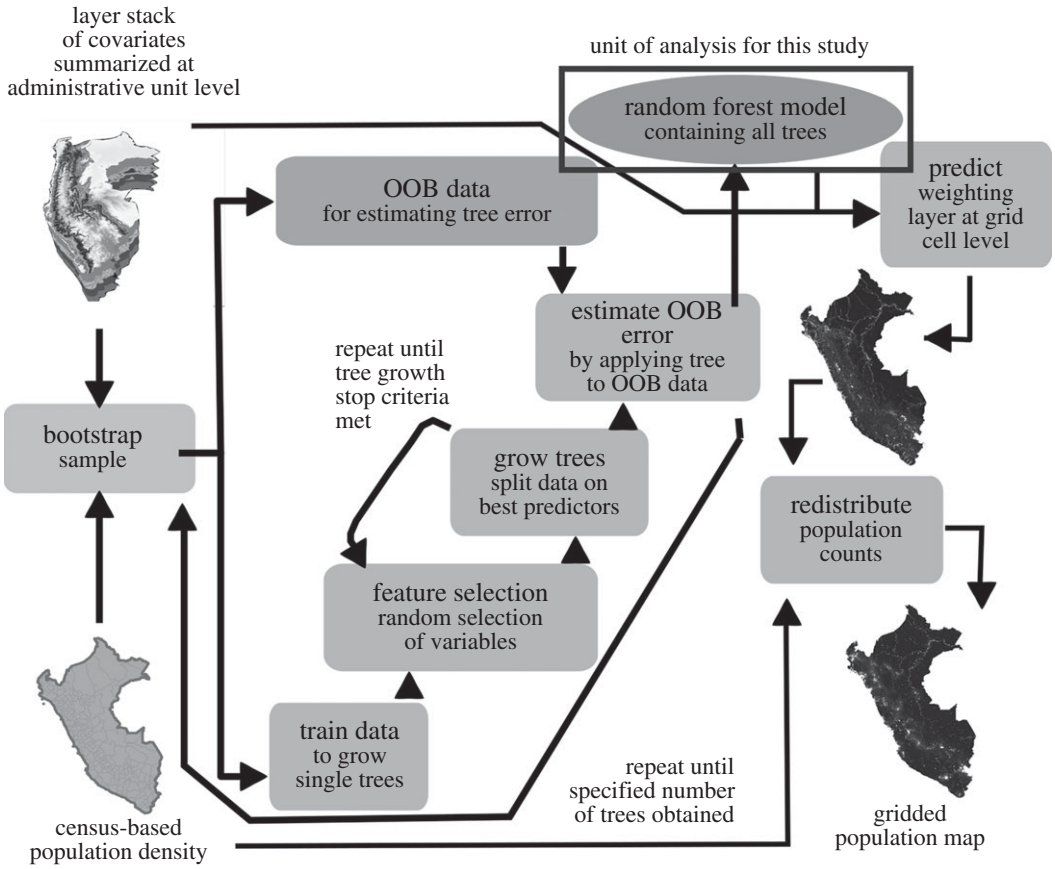

Figure 1. General process of using a random forest to created gridded population maps following Stevens *et al*. [26], where 'out-of-bag' (OOB) data are the approximately one-third of the data not sampled for training any single tree.

error of all trees [37]. The OOB error is also used for estimating covariate importance by randomly permutating a given covariate's OOB data with random noise and calculating the average per cent increase in the mean squared error, hereafter the Per.Inc.m.s.e., across all trees of the RF model which used the covariate [37]. For more details on the construction of RFs, see Breiman [37] and Liaw & Wiener [38].

The RF method outlined by Stevens *et al*. [26] uses an RF regression model and dasymetric mapping methods in a three-step process to estimate a population layer from input census and covariate data. The general steps are as follows: (i) iterative covariate selection for the RF model, (ii) the fitting of the RF model, using all available census units, and creation of a population density weighting layer from the created RF model, and (iii) the dasymetric redistribution of population counts from census-based administrative units to grid cells [29] using the population density weighting layer [26,32,33]. We give a general schematic of the RF process described by Stevens *et al*. [26] in figure 1. The covariate selection process is identical to step 2, but iterates until the removal of all covariates with a Per.Inc.m.s.e. less than zero. Data input to an RF model varies on a country-by-country basis with high-resolution country-specific datasets being used over coarser resolution default datasets, when available. This last detail required the standardization of what each covariate more generally represented to facilitate comparison across models.

2.2. Census data

For this investigation, we sampled countries ($n = 32$) from low- and middle-income countries in four regions of the world where available boundary-matched census data were available at an average spatial resolution (ASR) of 100 km^2 or below: Africa, Central America and the Caribbean (C. America and the Caribbean), South America (S. America) and Southeast Asia (S.E. Asia) [41]. The sampled countries, shown in figure 2, were modelled upon census data from varying years, with differing ASRs [41] of administrative units, and people per administrative unit, shown in table 1. These regions were selected because of their continued and rapidly growing importance in relation to world population [15,17].

2.3. Geospatial covariates and standardization

Human population density is highly correlated with environmental and physical factors [35], which can influence distributions of population. As indicated by the literature and availability of global data, the following factors were identified and used as predictive covariates: intensity of night-time lights [42], energy productivity of plants [43], topographic elevation and slope [44,45], climatic factors [46], type of land cover (LC) [27] and presence/absence of roads [47], water features [48], human settlements and urban areas [49], protected areas [50] and locations of points of interest (POIs) and facilities such as health centres and schools [51]. Rather than attempt to standardize the input covariates between countries, we used the most contemporary available datasets on a country-by-country basis to produce the population maps. See Stevens and co-workers [26] and [29] for a typical set of ancillary data included in a given model, with further details provided in Lloyd *et al*. [52].

For every model run, information about the RF model settings, covariates and their importance, metadata on the covariate datasets themselves and the general results of the RF model were output to summary files, which are included in the electronic supplementary material. From those summaries, we extracted the region modelled, the total variance explained by the model, the covariate names and the Per.Inc.m.s.e. for every covariate included in the model [37]. We then examined the covariates for all sampled countries to reclassify them into the covariate classification groups shown in table 2 as informed by common themes through the literature and patterns seen through population modelling of numerous countries. The primary

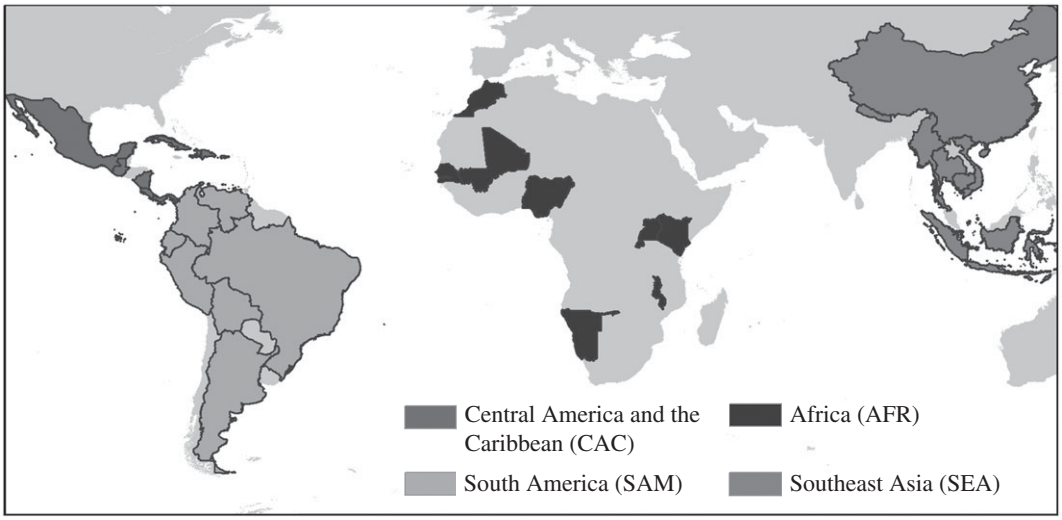

Figure 2. Countries for which boundary-matched census data were used in this study, from Africa, Central America and the Caribbean, South America and Southeast Asia.

purpose of this classification system was to facilitate comparisons between the country models via a standardized framework.

We would expect that covariates within the urban/suburban extents and built environment and urban/suburban proxy classes would be the most important for predicting population density as these typically capture settlements either implicitly or explicitly [10,64–66]. Transportation networks and facilities and service classes would also be expected to be consistently important as transportation networks exist solely to facilitate the movement of people, goods and ideas [66]. Responding to the classic 'location–allocation' problem, facilities and services, e.g. schools and health centres, are often located to promote access by and service to a population. Rivers/waterbodies/waterways are unique in that they can be used by people as both a transportation network and a resource, an attraction for population, but, in some cases, could be perceived as more hazard than resource, e.g. floods, and would therefore serve as a disincentive for a population locating near them. Previous studies have shown that landcover classes can be used for predicting population density by predicting either their absence, e.g. natural or bare surface land cover, or their presence due to their direct impact on land use (LU), e.g. cultivated land cover [8,27,32].

2.4. Analysis

From the independently modelled countries, we synthesized generalized data on the relative importance of various covariates in predicting population densities. All analysis and data handling was performed in the R Statistical Environment, version 3.2.2, with $\alpha = 0.05$ significance levels and appropriate corrections for multiple outcomes where indicated [67].

To account for the differing number of total covariates in each country's model, we calculated a weighted importance rank (WIR). Within each country, we ranked covariates by descending Per.Inc.m.s.e. and then weighted them by the total number of covariates in the final model for a given country, calculated as

$$\mathrm{WIR} = \frac{\text{within-country ranked importance}}{\text{total number of covariates in country model}}.$$

Within a given country, a WIR of zero indicates the covariate of highest importance and a WIR of 1 is the least important covariate. Hereafter, unless explicitly stated, within the text, variable class importance is referring to the WIR. To examine potential differences in variable class importance, we used both analytical and graphical methods.

Given the non-normal nature of the covariate importance data, we used the non-parametric form of the Kruskal–Wallis test to test for significant differences between covariate classes across all countries [68]. The inter-regional analyses were of a hierarchical nature using data subsets of a given covariate class and using the region category as the grouping variables, but still using the Kruskal–Wallis test [68,69]. The intra-regional analyses subset the data to a given region and a given covariate class then used a Kruskal–Wallis test to determine whether significant differences in importance for the given covariate class existed between countries of the same region [68]. If any of the Kruskal–Wallis tests were significant they were followed up with *post hoc* Dunn tests, using Holm's correction for multiple outcomes, to determine between which covariate classes or regions the significant differences occurred [70,71].

3. Results

The consistent patterns of covariate importance to predicting population density were observed between all sampled countries globally, with similar patterns observed between regions of countries. The correlates pertaining to urban areas and, more surprisingly, topographical features were the most important predictors of population density at all scales of analysis and were the only covariate categories which were consistently significantly more important than other categories, again at all scales.

3.1. Global

We present global covariate importances in figure 3. The five most important covariate classes, in descending order of median importance, were urban/suburban extents (0.32), built environment and urban/suburban proxies (0.35), climatic/environmental variables (0.37), populated place covariates (0.42) and transportation networks (0.50). This result matches expectations, as the five most important covariate classes (figure 3) are also the most often included in the final population models.

Globally, for predicting population density, we found that built environment covariates were significantly more important than classified populated place ($p < 0.01$), natural/semi-natural vegetation LC ($p < 0.01$), general classified LU ($p = 0.04$), protected LU ($p < 0.01$) and rivers/waterbodies/waterways covariates ($p < 0.01$). We also found that urban/suburban extents were significantly more important

Table 1. Sampled countries and selected characteristics including the variance explained by the country-specific random forest model. admin., administrative; avg., average.

country	ISO	region	census year (admin. level)	admin. units	avg. spatial resolution (km^2)	people per unit (thousands)	variance explained
Kenya	KEN	Africa	1999 (5)	6606	9	4.3	83%
Morocco	MAR	Africa	2004 (4)	1497	16	21	80%
Mali	MLI	Africa	2009 (4)	687	43	22	85%
Malawi	MWI	Africa	2008 (2)	12 557	22	59	79%
Namibia	NAM	Africa	2011 (2)	5475	12.28	21	96%
Nigeria	NGA	Africa	2006 (2)	774	34	205	88%
Rwanda	RWA	Africa	2002 (4)	9183	1.68	1.2	69%
Senegal	SEN	Africa	2009 (4)	331	24	37	91%
Uganda	UGA	Africa	2002 (4)	5018	7	6	85%
Bolivia	BOL	C. America and Caribbean	2012 (2)	112	97.7	91	65%
Costa Rica	CRI	C. America and Caribbean	2011 (3)	469	10.4	9.8	92%
Cuba	CUB	C. America and Caribbean	2012 (2)	168	25.6	68	82%
Dominican Republic	DOM	C. America and Caribbean	2010 (3)	155	17.6	64	86%
Guatemala	GTM	C. America and Caribbean	2012 (2)	333	18.0	46	80%
Haiti	HTI	C. America and Caribbean	2009 (4)	570	6.9	17	84%
Mexico	MEX	C. America and Caribbean	2010 (2)	2456	28.0	48	92%
Nicaragua	NIC	C. America and Caribbean	2012 (3)	137	29.4	43	79%
Panama	PAN	C. America and Caribbean	2010 (2)	74	31.04	49	74%
Puerto Rico	PRI	C. America and Caribbean	2010 (1)	78	13.3	48	74%
Argentina	ARG	S. America	2010 (2)	526	73.0	78	88%
Brazil	BRA	S. America	2010 (4)	5565	5.1	36	84%
Colombia	COL	S. America	2013 (4)	1115	32.0	42	84%
Ecuador	ECU	S. America	2010 (4)	978	16.2	15	82%
Peru	PER	S. America	2012 (2)	194	81.7	155	63%
Venezuela	VEN	S. America	2011 (2)	339	51.6	87	71%
Cambodia	KHM	S.E. Asia	2008 (3)	1621	10.51	8.6	92%
China	CHN	S.E. Asia	2010 (4)	2922	57.28	458	95%
Indonesia	IND	S.E. Asia	2010 (4)	79 277	4.91	3.0	81%
Myanmar	MMR	S.E. Asia	2014 (3)	326	45.29	164	94%
Nepal	NEP	S.E. Asia	2011 (4)	3973	6.08	6.8	92%
Thailand	THA	S.E. Asia	2010 (3)	7416	23.67	9.0	88%
Vietnam	VNM	S.E. Asia	2010 (3)	688	21.85	123	93%

than protected LU ($p < 0.01$). Furthermore, we observed that climatic/environmental variables were significantly more important than populated place ($p < 0.01$), natural/semi-natural vegetation LC ($p < 0.01$), general classified LU ($p = 0.02$), protected LU ($p < 0.01$) and rivers/waterbodies/waterways covariates ($p < 0.01$). Interestingly, we observed no

Table 2. Reclassification scheme to standardize covariates into variable classes representing spatial drivers and determinants of population. LC, thematically classified land cover; LU, classified land use; nat., natural; OSM, Open Street Map; semi.-nat., semi-natural; veg., vegetation. Note: The references are not exhaustive, but are characteristic of most models. Any of these covariates could be replaced by a country-specific dataset sourced from a one-off source or country partner. Refer to country-specific metadata files provided with the source download from www.worldpop.org.

aggregated variable class	drivers, correlates and covariates
natural/semi-natural vegetation land cover	LC nat. and semi-nat. veg.—woody [53,54]
	LC nat. and semi-nat. veg.—shrubs [53,54]
	LC nat. and semi-nat. veg.—herbaceous [53,54]
	LC nat. and semi-nat. veg.—other mix [53,54]
	LC nat. and semi-nat. veg.—aquatic veg. [53,54]
cultivated/managed land cover	LC cultivated terrestrial and managed lands [53,54]
natural bare surfaces land cover	LC natural bare surface [53,54]
artificial surface land cover	LC urban areas [53,54]
	LC rural settlement [53,54]
no data	LC no data [53,54]
residential land use	LU residential [55]
non-residential land use	LU industrial [55]
	LU farms [55]
protected land use	e.g. protected natural areas [56]
general classified land use	e.g. multiple classified land uses provided to model as a single covariate [55]
urban/suburban extents	global human settlement layer [57]
	Schneider MODIS [58]
built environment and urban/suburban proxies	LC urban areas + LC rural settlement [53]
	lights at night imagery [59]
	building footprints [55]
classified populated place (hierarchical)	e.g. city, town, village, etc. [55]
transportation networks	roads [55,60]
	railways [55]
climatic/environmental	elevation and slope [61]
	net primary productivity [62]
	temperature [63]
	precipitation [63]
facilities and services	schools [55]
	police [55]
	nutrition [55]
	health facilities [55]
places and POIs	OSM places [55]
	OSM POIs [55]
rivers/waterbodies/waterways	LC water [53,54]
	rivers [55]
	waterbodies/waterways [55,60]
populated place	e.g. gazetteer-type data [55,60]

significant difference in importance between the urban/suburban extents and the built environment and urban/suburban proxy classes. In table 3, we show test results for significant differences between covariates of the top five important covariate classes when compared with all other covariate classes. The complete results are detailed in the electronic supplementary material.

3.2. Inter-regional

Another POI was that the strong patterns of association seen at the global level were largely consistent when drivers and correlates were examined between regions. The only significant differences between regions were seen for the non-residential LU variable and the rivers/waterways/waterbodies variable, the latter shown in table 4. Non-residential

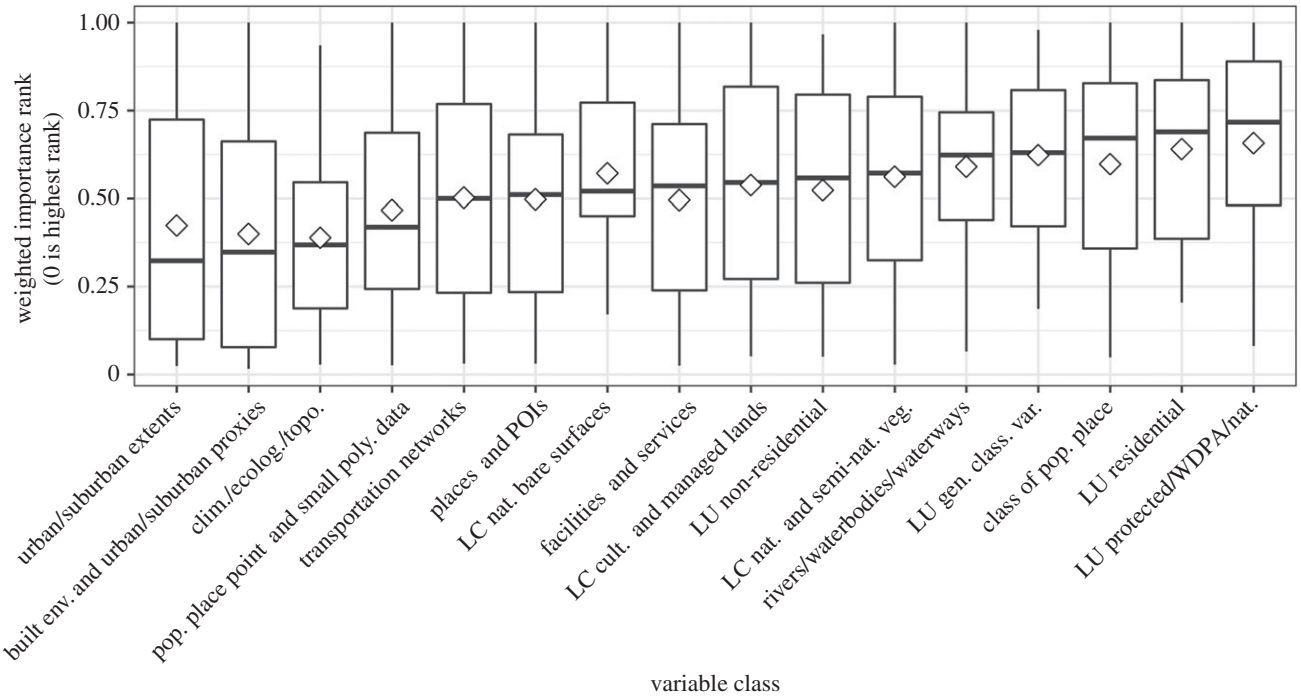

Figure 3. Global variable class weighted rank of importance based upon covariates included in a given country's final model, where zero represents the highest rank. The mean is represented by a white diamond; the median is represented by the black bar; and the whiskers represent the maximum and minimum values within 1.5× the inter-quartile range. See table 2 for descriptions and references for the variable classes. LC, land cover; LU, land use; WDPA, World Database on Protected Areas.

Table 3. Selected results of the pairwise *post hoc* Dunn test with Holm's correction for multiple outcomes of global WIR of covariate classes. See table 2 for descriptions and references for the variable classes. LC, land cover; LU, land use. See the electronic supplementary material for results across all classes. Global Kruskal–Wallis results: d.f. = 15, chi-squared = 96.147, $p < 0.01$. Full precision of the values is provided in the electronic supplementary material.

	corrected *Z*-value (corrected *p*-values)*				
variable class	built env. and urban/ suburb. proxies	climatic/ environmental	populated place	transportation networks	urban/suburb. extents
class of pop. place	5.04 (<0.01)	5.53 (<0.01)	2.41 (1.00)	2.41 (1.00)	3.43 (0.06)
climatic/environmental	0.30 (1.00)	—	1.49 (1.00)	3.20 (0.14)	0.72 (1.00)
facilities and services	2.06 (1.00)	2.36 (1.00)	0.48 (1.00)	0.16 (1.00)	1.27 (1.00)
cultivated/managed LC	3.43 (0.37)	3.20 (0.14)	1.18 (1.00)	0.74 (1.00)	1.98 (1.00)
natural/semi-natural vegetation LC	4.82 (<0.01)	5.44 (<0.01)	1.90 (1.00)	1.76 (1.00)	2.98 (0.28)
nat. bare surfaces LC	3.19 (0.14)	3.46 (0.06)	1.60 (1.00)	1.27 (1.00)	2.35 (1.00)
general classified LU	3.58 (0.04)	3.81 (0.02)	2.15 (1.00)	1.93 (1.00)	2.84 (0.42)
non-residential LU	1.55 (1.00)	1.71 (1.00)	0.64 (1.00)	0.25 (1.00)	1.16 (1.00)
protected LU	5.52 (<0.01)	5.91 (<0.01)	3.19 (0.14)	3.31 (0.10)	4.13 (<0.01)
residential LU	3.37 (0.08)	3.56 (0.04)	2.16 (1.00)	1.93 (1.00)	2.77 (0.52)
places and POIs	2.08 (1.00)	2.38 (1.00)	0.51 (1.00)	0.11 (1.00)	1.29 (1.00)
populated place	1.26 (1.00)	1.49 (1.00)	—	0.68 (1.00)	0.69 (1.00)
rivers/waterbodies/ waterways	4.80 (<0.01)	5.28 (<0.01)	2.27 (1.00)	2.20 (1.00)	3.27 (0.11)
transportation networks	2.76 (0.52)	3.20 (0.14)	0.68 (1.00)	—	1.61 (1.00)
urban/suburban extents	0.48 (1.00)	0.72 (1.00)	0.69 (1.00)	1.61 (1.00)	—

LU was significantly more important in C. America and the Caribbean than in S. America ($p = 0.02$; $Z = 2.35$). As shown in table 4, rivers/waterbodies/waterways were significantly more important in Africa ($p < 0.01$; $Z = 3.78$) and S.E. Asia ($p < 0.01$; $Z = 4.08$) than in C. America and the Caribbean.

The consistency of importances within covariate classes across regions becomes apparent when plotting the

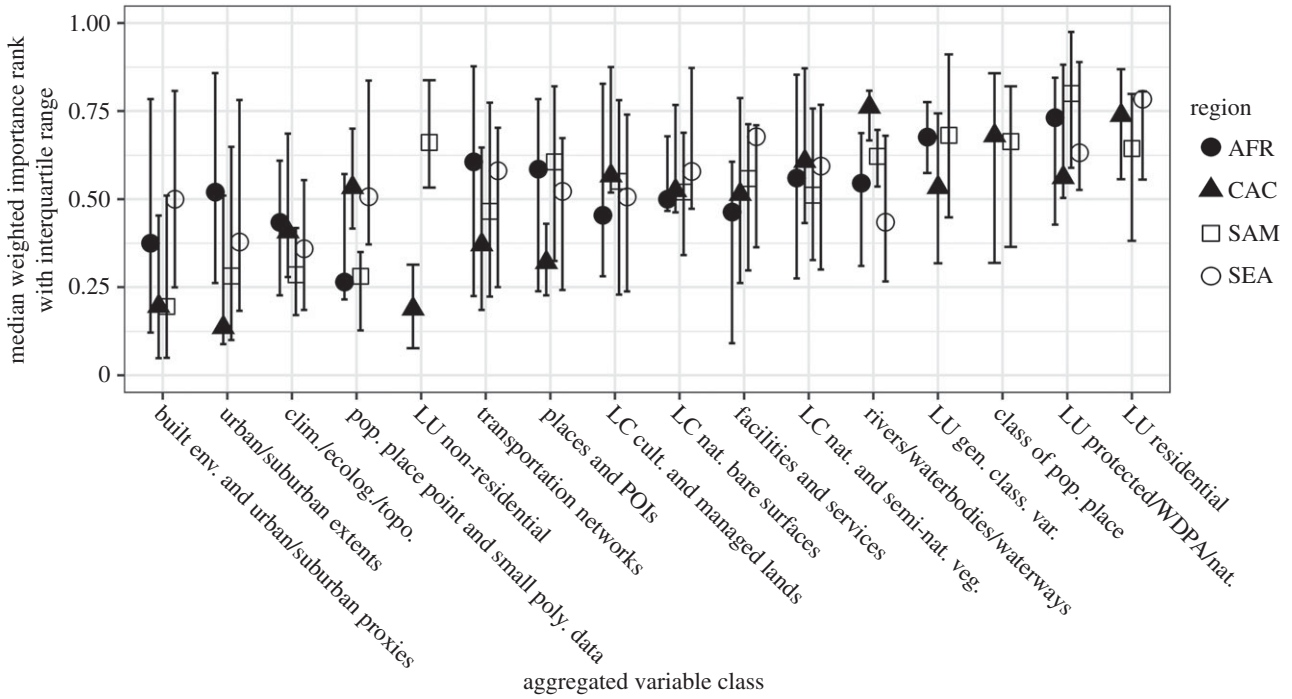

Figure 4. Regional line and dot plot of variable class WIR with the median marked by the dot and the inter-quartile range demarcated by brackets. Note that not all regions have all variable classes. See table 2 for descriptions and references for the variable classes.

Table 4. Results of the pairwise Dunn test with Holm's correction for differences in WIR of variable class by region within the rivers/waterbodies/waterways class. Corrected *Z*-score and corrected *p*-value, in parentheses, are given. Full results for all variable classes between regions, including non-significant findings, are provided in the electronic supplementary material. Kruskal–Wallis results: d.f. = 3, chi-squared = 20.281, $p < 0.01$.

region	Africa	C. America and Caribbean	S. America
C. America and Caribbean	3.78 (<0.01)	—	—
S. America	1.21 (0.45)	2.32 (0.08)	—
S.E. Asia	0.77 (0.45)	4.08 (<0.01)	1.79 (0.22)

importance, with the inter-quartile range (IQR), as done in figure 4. It can first be noted that many of the covariate class IQRs overlap between regions, with very similar median importances and variation seen for climatic/environmental covariates, transportation networks and cultivated/managed LC. There is more variation in importance than expected between regions for covariates of urban/suburban extents and the built environment and urban/suburban proxies. The findings from table 4, and all the inter-regional tests included in the electronic supplementary material, agree with the distributions shown in figure 4.

3.3. Intra-regional

Like global patterns, there were no differences between the importance of the covariates urban/suburban extents and built environment and urban/suburban proxies within any region. Within any single region, we found no significant differences in patterns of importance between countries for all given covariate classes. However, between covariate classes across all countries within a given region, we found significant differences within the C. America and the Caribbean and S. America regions and display these in table 5. Similar to the global results, we found within S. America that built environment and urban/suburban proxies were significantly more important than classified populated place ($p < 0.01$), protected LU ($p < 0.01$) and rivers/waterbodies/waterways covariates ($p = 0.01$). Also within S. America, we found that climatic/environmental variables were significantly more important than classified populated place ($p < 0.01$), natural/semi-natural vegetation LC ($p = 0.02$), general classified LU ($p = 0.04$), protected LU ($p < 0.01$) and rivers/waterbodies/waterways covariates ($p < 0.01$). For C. America and the Caribbean, we found that the covariates regarding built environment and urban/suburban proxies ($p < 0.01$), transportation networks ($p = 0.03$), urban/suburban extents ($p < 0.01$) and climatic/environmental variables ($p = 0.02$) were significantly more important than rivers/waterbodies/waterways covariates. Additionally, built environment and urban/suburban proxies were found to be significantly more important than classified populated place ($p = 0.01$), natural/semi-natural vegetation LC ($p < 0.01$) and protected LU ($p < 0.05$). Full results including the non-significant findings are included in the electronic supplementary material. We illustrate the consistency of the importance of distribution and their relative importance regionally for each covariate class graphically in figure 5.

4. Discussion

The majority of predicted population growth across the globe by 2050 is expected to occur in low- and middle-income countries [14,15,17]. With this predicted growth in population and urbanization challenges are expected to arise regarding food security, health and infrastructure, to name but a few

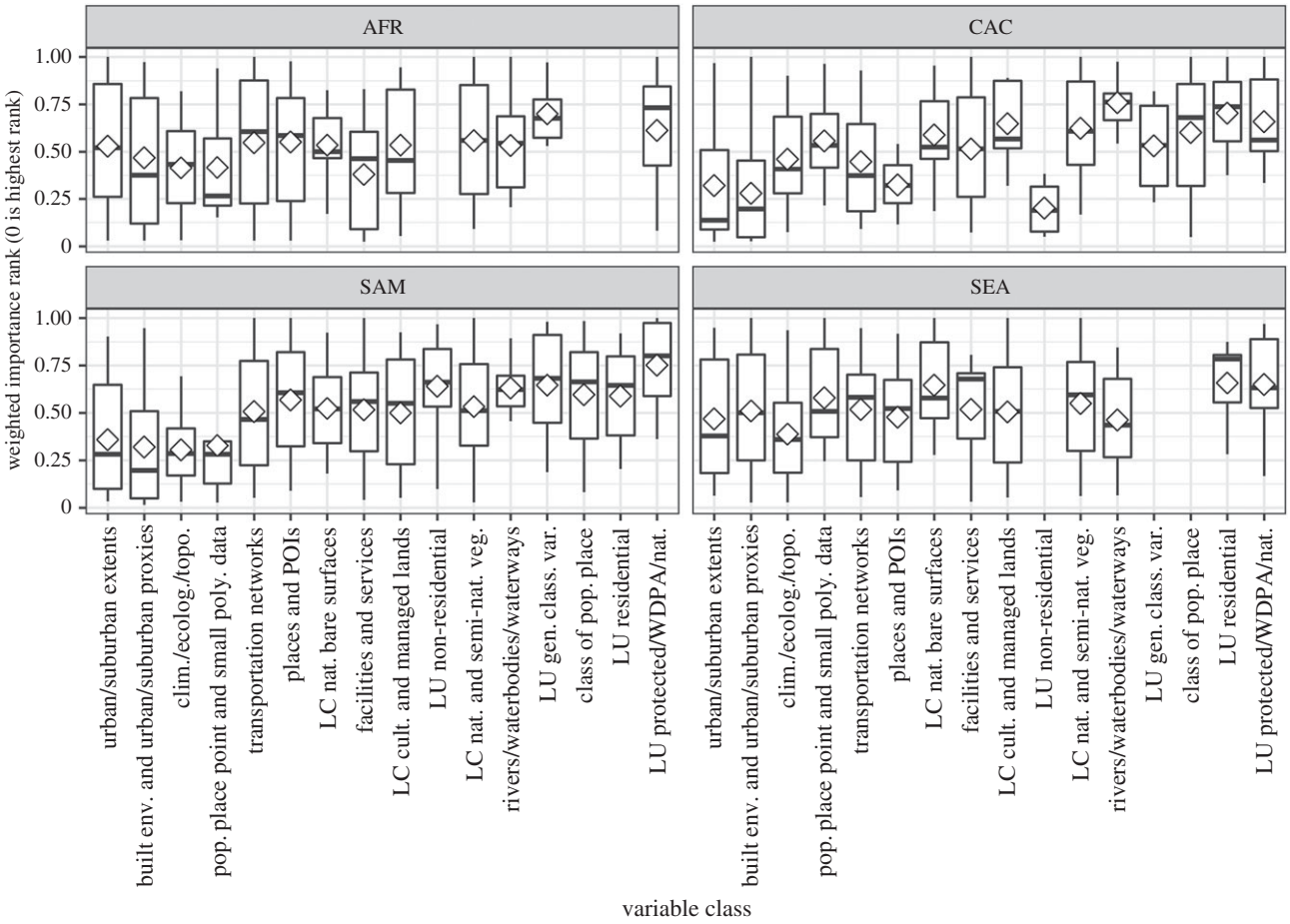

Figure 5. Regional variable class weighted rank of importance based upon covariates included in a given country's final model, where zero represents the highest rank. The mean is represented by a white diamond; the median is represented by the black bar; and the whiskers represent the maximum and minimum values within 1.5× the inter-quartile range. See table 2 for descriptions and references for the variable classes.

Table 5. Selected results of the pairwise Dunn test with Holm's correction for differences in WIR by region between variable classes. Corrected *Z*-scores and corrected *p*-values, in parentheses, are given. Full results between all variable classes within regions, including non-significant findings, are provided in the electronic supplementary material. Full precisions of values are provided in the electronic supplementary material.

region	variable class	built env. and urban/ suburban proxies	climatic/ environmental	urban/suburban extents	transportation networks	populated place
S. America	classified populated place	4.54 (<0.01)	5.09 (<0.01)	2.73 (0.63)	1.69 (1.00)	2.76 (0.57)
	natural/semi-natural vegetation LC	3.73 (0.10)	3.73 (0.02)	1.94 (1.00)	0.46 (1.00)	2.06 (1.00)
	general classified LU	3.34 (0.09)	3.56 (0.04)	2.48 (1.00)	1.50 (1.00)	2.57 (0.95)
	protected LU	4.29 (<0.01)	4.52 (<0.01)	3.32 (0.10)	2.55 (1.00)	3.36 (0.08)
	rivers/waterbodies/ waterways	3.82 (0.01)	4.14 (<0.01)	2.65 (0.77)	1.63 (1.00)	2.72 (0.63)
C. America and Caribbean	classified populated place	3.85 (0.01)	1.76 (1.00)	2.63 (0.88)	1.84 (1.00)	0.39 (1.00)
	natural/semi-natural vegetation LC	4.62 (<0.01)	2.30 (1.00)	3.03 (0.26)	2.36 (1.00)	0.61 (1.00)
	protected LU	3.52 (<0.05)	1.88 (1.00)	2.66 (0.81)	1.95 (1.00)	0.75 (1.00)
	rivers/waterbodies/ waterways	5.66 (<0.01)	3.66 (0.03)	4.07 (<0.01)	3.69 (0.03)	1.75 (1.00)

[72–76]. These continued and heightened concerns regarding the implications of the rapid pace of shifting populations in low- and middle-income countries ensure a continued demand for high-resolution gridded population maps in these regions of the world. This continued demand reinforces why understanding the drivers of the spatial distribution of

populations to improve population mapping is important. Moreover, an improved understanding of the fundamental drivers of population distributions and their spatial variations is of value for modelling future growth and designing strategies around such models.

Our results show that variables related to built/urban areas and to climatic/environmental covariates were the most important for predicting population density and were the only covariate classes that were significantly more important than other variable classes, regardless of the scale of analysis. This study begins to quantify commonly held concepts regarding the drivers and correlates of human population distributions, e.g. urban areas are associated with denser populations. Having quantified these patterns globally and regionally allows future work on the more unique aspects of location-specific distributional relationships of populations to be placed within the context of these larger-scale findings, and to help relate observed and past population distributions to historical and cultural contexts and the presence or absence of resources/hazards.

The finding that built area-related covariates were the most important in predicting population density should not be a surprise and it aligns with expectations that an estimated 54% of the world's population live in urbanized areas [15]. There are numerous examples where population density was an important predictor of urban area extent [77–80]. This study shows that this relationship goes in the other direction as well with built area extent being important in predicting population density. However, caution should be used when using the newer urban/settlement feature datasets such as global human settlement layer and global urban footprint. While they are improvements on the thematically classified 'urban', making use of spectrally and spatially refined optical and radar-based data, they are known to be most accurate in dense urbanized areas [64,65], leading to population model biases in less densely populated or rural contexts by virtue of the settlements being missed in the input covariates [26].

We were surprised how important the climatic/environmental covariate category was in predicting population density. While the category was not broken up for subsequent testing, by examining the covariate importance plots of individual countries we believe that this importance was largely driven by elevation covariates, including derived slope. Previous studies have shown that population is prevalent in the lower elevations of resource-rich coastal zones, deltas and river valleys [81–83] and it is simply easier to build on relatively shallow–moderate slopes than on steep slopes. There is also precedence for transportation and elevation covariates being predictive of urban or built land cover, corroborated by our finding that transportation networks and climatic/environmental covariate classes were consistently important predictors of population density [27,84,85]. Water-related covariates being consistently less important than crop or natural vegetation landcover covariates (figure 3) could be a result of the resource/hazard relationship [86] that populations have with waterbodies, which of course is highly context dependent.

Differing data quality of input covariates to the models analysed here should be kept in mind when interpreting these results as they directly affect the observed importance, or non-importance, of the covariates. For instance, the significant difference seen between C. America and the Caribbean and Africa and between C. America and the Caribbean and S.E. Asia within the rivers/waterbodies/waterways covariate class (table 4) is most likely to be due to the different thematic land cover sources used for those regions. While all landcover data used were adjusted to a standard thematic framework and resampled to 100 m [27], the majority of the Africa models used the 300 m resolution Globcover data whereas the S.E. Asia and the C. America and the Caribbean data were based upon the commercial, 30 m resolution, Geocover data [28]. While C. America and the Caribbean and S.E. Asia both used the Geocover dataset, they also sourced OpenStreetMap [55] for data pertaining to river features. OpenStreetMap varies widely as to completeness, coverage and data quality [87,88]. So, we would speculate that the observed significant differences were not likely to be indicative of actual differences in how the population relates to water features between those regions, but are the result of different data sources for the built area-related covariates being used (table 2). Similar differing data quality or completeness issues are likely to be at the source of the significant differences between regions seen for the residential LU variable, which is entirely based on OpenStreetMap data [55].

These findings are valid only for a specific spatial resolution and modelling scale that may or may not maintain the same structures and relationships at a finer scale, as is typically the case with the modifiable areal unit problem (MAUP) [89]. All covariates are affected to some degree because they are all resampled to 100 m and are further aggregated by some summary measure at the administrative unit level prior to input in the RF from which our covariate importance metrics are derived [26]. Variations in data quality of the census-based population counts and the differing number of administrative units used in each region's countries modelled can partly explain the variance in importances within variable classes between regions. This follows the scale effect of the MAUP, which states that as the number of areal units is decreased there is a decrease in the variability of the observations corresponding to the areal units [89]. The potential of the coarseness of the polygonal census units to have an effect on this variability is less clear, but is likely to have an effect similar to the MAUP zonation effect [89]. So, while we observed very consistent patterns of importance between classes of variables and population density, this is based upon country-level averages of importance derived from a country-specific level of sub-national units and then analysed at the country level across all countries and between and within regional groupings of countries. Were we to change the groupings, e.g. change the level of sub-national units from which a country-level RF is constructed, then, following the MAUP, the results would be likely to change. However, given that no significant differences in importance for any covariate class between countries within a given region were found, it would appear that the regional groupings maximized internal homogeneity, better facilitating inter-regional testing for differences.

There are inferential limits to using the RF model to identify/approximate the structure of covariate class relationships to population density. Unlike multiple linear regressions or single regression trees where coefficients and confidence intervals can be quantified, the numerous trees in an RF preclude the tracing of the regression from input to prediction [37]. The strength of an RF to capture nonlinear relationships of covariates and their complex interactions, through its numerous trees, does not make for simple interpretations of the underlying mechanisms of the modelled phenomenon, in this case the driver and correlates of population

distribution [37]. Covariate importance within an RF is also complex because of those same nonlinear relationships and interactions and results in a covariate's importance within an RF being highly conditional on all other covariates present, with similar results not guaranteed in other models, even for the same country [38].

Another consideration when evaluating the importance of covariate classes and their relationships to population density is the varying temporality of the covariate datasets, which may not match the date of the input census data. Therefore, the modelled relationships are imperfect to begin with, as it is impossible to have complete temporal agreement between all input datasets because of well-known availability constraints. Furthermore, the quality of census data varies from country to country as well as from census to census, with completeness and spatial resolution of the administrative units being variable.

Further investigating these covariates in relation to population density could involve utilizing a different modelling framework that would allow for more inferential power as to the structure and nature of the relationships between these covariates and population density. Additionally, focusing our study on specific covariate classes, such as the urban-/suburban-related variable classes, by sourcing novel and forthcoming datasets that help illuminate the heterogeneity within these areas, both internally and across different countries and regions, could increase the predictive ability of a population model regardless of the framework. As these population datasets are scaled up to global extent, the question occurs as to whether these trends persist in high-income regions and once a consistent set of covariates is used for modelling all countries.

Better mapping of potential trends regarding drought [90], water distribution [91], crop distribution [92] and forest distribution [93] continue to improve and refine our spatial awareness of resource distribution, change and environmental patterns, globally. The relationships between population distribution and various ancillary datasets outlined in this paper provide relevant information for future work examining how populations may react to a continually changing landscape. In addition, potential exists to integrate such temporally dynamic datasets into gridded population models for better informing population distribution, not only over space but also over time [94]. However, this study is simply a cross section of covariate relationships to population density; a key question is whether these relationships remain static or are dynamic through time and the answer to that question is of great importance to population growth models, and other population-related fields, looking backwards and forwards through time.

Data accessibility. All original extracted data and code used for analysis has been provided in the electronic supplementary material. The large size of the original model objects precluded their inclusion in the electronic supplementary material. Requests for the original model objects should be sent to the corresponding author.

Authors' contributions. J.J.N. was responsible for research design, coding, data collection, management, processing statistical analyses, interpretation and drafting and production of the final manuscript. F.R.S. was responsible for research design, coding, data collection, interpretation and production of the final manuscript. A.E.G. was responsible for research design, data collection, interpretation and production of the final manuscript. C.L. was responsible for data collection and interpretation. A.S. was responsible for data collection and interpretation, and production of the final manuscript. G.H. was responsible for data collection. N.N.P. was responsible for data collection. A.J.T. was responsible for overall scientific management, interpretation and drafting and production of the final manuscript. All authors gave final approval for publication.

Competing interests. The authors declare that they have no competing interests.

Funding. J.J.N., F.R.S. and A.E.G. are supported by funding from Google. J.J.N. is supported by a grant from the Economic and Social Research Council Southcoast Doctoral Training Programme. A.S. is supported by funding from the Bill & Melinda Gates Foundation (OPP1106427, OPP1032350). A.J.T. is supported by funding from NIH/NIAID (U19AI089674), the Bill & Melinda Gates Foundation (OPP1106427, OPP1032350, OPP1134076), the Clinton Health Access Initiative, the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security and the Fogarty International Center, National Institutes of Health, and a Wellcome Trust Sustaining Health Grant (106866/Z/15/Z). This work forms part of the outputs of the WorldPop Project (www.worldpop.org) and the Flowminder Foundation (www.flowminder.org). The funders had no role in study design, data collection and analysis, decision to publish and preparation of the manuscript.

Acknowledgements. We would like to thank Amy Ninneman for her copy-editing skills, which she graciously volunteered to the writing of this manuscript. The rest of the WorldPop team are acknowledged for their support and feedback on preliminary findings. This work forms part of the WorldPop Project (www.worldpop.org.uk) and Flowminder Foundation (www.flowminder.org).

References

1. McGranahan G, Balk D, Anderson B. 2007 The rising tide: assessing the risks of climate change and human settlements in low elevation coastal zones. *Environ. Urban* **19**, 17–37. (doi:10.1177/0956247807076960)
2. Kirch PV, Hartshorn AS, Chadwick OA, Vitousek PM, Sherrod DR, Coil J, Holm L, Sharp WD. 2004 Environment, agriculture, and settlement patterns in a marginal Polynesian landscape. *Proc. Natl Acad. Sci. USA* **101**, 9936–9941. (doi:10.1073/pnas.0403470101)
3. Parsons JR. 1972 Archeological settlement patterns. *Annu. Rev. Anthropol.* **1**, 127–150. (doi:10.1146/annurev.an.01.100172.001015)
4. Trigger BG. 1967 Settlement archaeology. Its goals and promises. *Am. Antiq.* **32**, 149–160. (doi:10.2307/277900)
5. Weisler M, Kirch PV. 1985 The structure of settlement space in a Polynesian chiefdom: Kawela, Moloka'i, Hawaiian Islands. *New Zeal J. Archaeol.* **7**, 129–158.
6. Trigger BG. 1971 Archaeology and ecology. *World Archaeol.* **2**, 321–336. (doi:10.1080/00438243.1971.9979483)
7. Sanderson EW, Jaiteh M, Levy MA, Redford KH, Wannebo AV, Woolmer G. 2002 The human footprint and the last of the wild. *Bioscience* **52**, 891–904. (doi:10.1641/0006-3568(2002)052[0891:THFATL]2.0.CO;2)
8. Meyer WB, Turner BL. 1992 Human population growth and land-use/cover change. *Annu. Rev. Ecol. Syst.* **23**, 39–61. (doi:10.1146/annurev.es.23.110192.000351)
9. Verburg PH, Schot PP, Dijst MJ, Veldkamp A. 2004 Landuse change modelling: current practice and research priorities. *GeoJournal* **61**, 309–324. (doi:10.1007/s10708-004-4946-y)
10. Elvidge CD, Imhoff ML, Baugh E, Hobson VR, Nelson I, Safran J, Dietz JB, Tuttie BT. 2001 Night-time lights of the world: 1994–1995. *Photogramm. Remote Sens.* **56**, 81–99. (doi:10.1016/S0924-2716(01)00040-5)
11. Pozzi F, Small C. 2005 Analysis of urban land cover and population density in the United States.

Photogramm. Eng. Remote Sens. **71**, 719–726. (doi:10.14358/PERS.71.6.719)

12. Holdren JP, Ehrlich D. 1974 Human population and the global environment: population growth, rising per capita material consumption, and disruptive technologies have made civilization a global ecological force. *Am. Sci.* **62**, 282–292.
13. Harte J. 2007 Human population as a dynamic factor in environmental degradation. *Popul. Environ.* **28**, 223–236. (doi:10.1007/s11111-007-0048-3)
14. Cohen JE. 2003 Human population: the next half century. *Science* **302**, 1172–1175. (doi:10.1126/science.1088665)
15. UN. 2015 *World urbanization prospects: the 2014 revision*. New York, NY: UN.
16. UN. 2015 *World population prospects: the 2015 revision—Key findings*. Washington, DC: UN.
17. UN. 2015 *World population prospects: the 2014 revision*. Washington, DC: UN.
18. Balk D, Yetman G. 2004 *The global distribution of population: evaluating the gains in resolution refinement*. New York, NY: Center for International Earth Science Information Network, Columbia University.
19. UNEP. 2004 UNEP/GRID—Sioux Falls Clearinghouse. United Nations Environment Programme. See https://na.unep.net/siouxfalls/datasets/datalist.php.
20. CIESIN. 2011 *Global rural urban mapping project* (*GRUMP*). Palisades, NY: Center for International Earth Science Information Network.
21. Balk D, Deichmann U, Yetman G, Pozzi F, Hay S, Nelson A. 2006 Determining global population distributions: methods, applications, and data. *Adv. Parasitol.* **62**, 119–156. (doi:10.1016/S0065-308X(05)62004-0)
22. Dobson JE, Bright EA, Coleman PR, Durfee RC. 2000 A global population database for estimating populations at risk. *Photogramm. Eng. Remote Sens.* **66**, 849–857.
23. Bhaduri B, Bright E, Coleman P. 2007 Landscan USA: a high resolution geospatial and temporal modeling approach for population distribtion and dynamics. *GeoJournal* **69**, 103–177. (doi:10.1007/s10708-007-9105-9)
24. Cheriyadat A, Bright E, Potere D, Bhaduri B. 2007 Mapping of settlements in high-resolution satellite imagery using high performance computing. *GeoJournal* **69**, 119–129. (doi:10.1007/s10708-007-9101-0)
25. European Commission, Columbia University for IESIN-C. 2015 *GHS population grid, derived from GPW4, multitemporal [1975, 1990, 2000, 2015]*. Brussels, Belgium: European Commission, Joint Research Centre. See http://data.europa.eu/89h/jrc-ghslghs_pop_gpw4_globe_r2015a.
26. Stevens FR, Gaughan AE, Linard C, Tatem AJ. 2015 Disaggregating census data for population mapping using random forests with remotely-sensed data and ancillary data. *PLoS ONE* **10**, e0107042. (doi:10.1371/journal.pone.0107042)
27. Linard C, Gilbert M, Tatem AJ. 2011 Assessing the use of global land cover data for guiding large area population distribution modelling. *GeoJournal* **76**, 525–538. (doi:10.1007/s10708-010-9364-8)
28. Gaughan AE, Stevens FR, Linard C, Jia P, Tatem AJ. 2013 High resolution population distribution maps for Southeast Asia in 2010 and 2015. *PLoS ONE* **8**, e55882. (doi:10.1371/journal.pone.0055882)
29. Sorichetta A, Hornby GM, Stevens FR, Gaughan AE, Linard C, Tatem AJ. 2015 High-resolution gridded population distribution datasets of Latin America in 2010, 2015, and 2020. *Sci. Data* **2**, 150045. (doi:10.1038/sdata.2015.45)
30. Leyk S, Ruther M, Buttenfield BP, Nagle NN, Stum AK. 2014 Modeling residential developed land in rural areas: a size-restricted approach using parcel data. *Appl. Geogr.* **47**, 33–45. (doi:10.1016/j.apgeog.2013.11.013)
31. Irwin EG, Cho H, Bockstael NE. 2007 Measuring the amount and pattern of land development in nonurban areas. *Rev. Agric. Econ.* **29**, 494–502. (doi:10.1111/j.1467-9353.2007.00360.x)
32. Mennis J, Hultgren T. 2006 Intelligent dasymetric mapping and its application to areal interpolation. *Cartogr. Geogr. Inf. Sci.* **33**, 179–194. (doi:10.1559/152304006779077309)
33. Mennis J. 2003 Generating surface models of population using dasymetric mapping. *Prof. Geogr.* **55**, 31–42.
34. Tapp AF. 2010 Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartogr. Geogr. Inf. Sci.* **37**, 215–228. (doi:10.1559/152304010792194976)
35. Nagle NN, Buttenfield BP, Leyk S, Spielman S. 2014 Dasymetric modeling and uncertainty. *Ann. Assoc. Am. Geogr.* **104**, 80–94. (doi:10.1080/00045608.2013.843439)
36. Farror DE, Glauber RR. 1967 Multicolinearity in regression analysis: the problem revisited. *Rev. Econom. Stat.* **56**, 92–107. (doi:10.2307/1937887)
37. Breiman L. 2001 Random forests. *Mach. Learn.* **45**, 5–32. (doi:10.1023/A:1010933404324)
38. Liaw A, Wiener M. 2002 Classification and regression by RandomForest. *R News* **3**, 18–22.
39. Rodriguez-Galiano VF, Ghimire B, Rogan J, Chica-Olmo M, Rigol-Sanchez JP. 2012 An assessment of the effectiveness of a random forest classifier for landcover detection. *Photogramm. Remote Sens.* **67**, 93–104. (doi:10.1016/j.isprsjprs.2011.11.002)
40. Breiman L. 1996 Bagging predictors. *Mach. Learn.* **24**, 123–140. (doi:10.1007/BF00058655)
41. Tobler W, Deichmann U, Gottsegen J, Maloy K. 1997 World population in a grid of spherical quadrilaterals. *Int. J. Popul. Geogr.* **3**, 203–225. (doi:10.1002/(SICI)1099-1220(199709)3:3<203::AID-IJPG68>3.0.CO;2-C)
42. Briggs DJ, Gulliver J, Fecht D, Vienneau DM. 2007 Dasymetric modelling of small-area population distribution using land cover and light emissions data. *Remote Sens. Environ.* **108**, 451–466. (doi:10.1016/j.rse.2006.11.020)
43. Luck JW. 2007 The relationships between net primary productivity, human population density and species conservation. *J. Biogeogr.* **34**, 201–212. (doi:10.1111/j.1365-2699.2006.01575.x)
44. Cohen JE, Small C. 1998 Hypsographic demography: the distribution of human population density by altitude. *Proc. Natl Acad. Sci. USA* **95**, 14 009–14 014. (doi:10.1073/pnas.95.24.14009)
45. Schumacher JV, Redmond RL, Hart MM, Jensen ME. 2000 Mapping patterns of human use and potential resource conflict on public lands. *Environ. Monit. Assess.* **64**, 127–137. (doi:10.1023/A:1006496023729)
46. Small C, Cohen JE. 2004 Continental physiography, climate, and the global distribution of human population. *Curr. Anthropol.* **45**, 269–277. (doi:10.1086/382255)
47. Reibel M, Bufalino ME. 2005 Street-weighted interpolation techniques for demographic count estimation in incompatible zone systems. *Environ. Plan A* **37**, 127–139. (doi:10.1068/a36202)
48. Kummu M, de Moel H, Ward PJ, Varis O. 2011 How close do we live to water? A global analysis of human population. *PLoS ONE* **6**, e20578. (doi:10.1371/journal.pone.0020578)
49. Tatem AJ, Noor AM, von Hagen C, Di Gregorio A, Hay SI. 2007 High resolution population maps for low income nations: combining land cover and census in East Africa. *PLoS ONE* **2**, e1298. (doi:10.1371/journal.pone.0001298)
50. Luck JW. 2007 A review of the relationships between human population density and biodiversity. *Biol. Rev.* **82**, 607–645. (doi:10.1111/j.1469-185X.2007.00028.x)
51. Linard C, Tatem AJ, Stevens FR, Gaughan AE, Patel NN, Huang Z. 2014 Use of active and passive VGI data for population distribution modelling: experience from the WorldPop project. In *Proc. of the Eighth Int. Conf. on Geographic Information Science, Vienna, Austria, 24–26 September 2014,* pp. 1–16. Vienna, Austria: Vienna University of Technology.
52. Lloyd CT, Sorichetta A, Tatem AJ. 2017 High resolution global gridded data for use in population studies. *Sci. Data* **4**, 170001. (doi:10.1038/sdata.2017.1)
53. European Space Agency. 2013 *Globcover, version 2.3*. Paris, France: ESA.
54. MDA Federal. 2004 *Landsat geocover 2000 ETM edition mosaics, v. 1.0*. Sioux Falls, SD: USGS.
55. OpenStreetMap Contributers. 2017 *Openstreetmap* (*OSM*) *database*. Sutton Coldfield, UK: OSM.
56. Protected Planet. 2015 World database on protected areas, 2nd edn. IUCN & UNEP. See https://www.protectedplanet.net/.
57. Pesaresi M *et al.* 2016 *Operating procedure for the production of the global human settlement layer from landsat data of the epochs 1975, 1990, 2000, and 2014*. Luxembourg: Publications Office of the European Union.
58. Schneider A, Friedl MA, Potere D. 2010 Mapping urban areas using MODIS 500-m data: new methods and datasets based on 'urban ecoregions'. *Remote Sens. Environ.* **114**, 1733–1746. (doi:10.1016/j.rse.2010.03.003)

59. Earth Observation Group NNGDC. 2013 *VIIRS nighttime lights—2012* (*two month composite*). Boulder, CO: NOAA National Centers for Environmental Information.
60. United States National Imagery and Mapping Agency. 1997 *Vector map level 0* (*VMAP0*). Bethesda, MD: USGS Information Services.
61. Lehner B, Verdin K, Jarvis A. 2008 New global hydrography derived from spaceborne elevation data. *Eos, Trans. Am. Geophys. Union* **89**, 93–94. (doi:10.1029/2008EO100001)
62. Numerical Terradynamic Simulation Group/ University of Montana. 2015 *MODIS 17a3, v. 55*. Sioux Falls, SD: NASA LP DAAC, USGS Earth Resources Observation and Science (EROS) Center.
63. Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A. 2005 Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **25**, 1965–1978. (doi:10.1002/joc.1276)
64. Esch T *et al.* 2013 Urban footprint processor—fully automated processing chain generating settlement masks from global data of the TanDEM-X Mission. *IEEE Geosci. Remote Sens. Lett.* **10**, 1617–1621. (doi:10.1109/LGRS.2013.2272953)
65. Pesaresi M *et al.* 2013 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.* **6**, 2102–2131. (doi:10.1109/JSTARS.2013.2271445)
66. Zandbergen PA, Ignizio DA. 2010 Comparison of dasymetric mapping techniques for small-area population estimates. *Cartogr. Geogr. Inf. Sci.* **37**, 199–214. (doi:10.1559/152304010792194985)
67. Team Rc. 2015 *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
68. Kruskal WH, Wallis WA. 1952 Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* **47**, 583–621. (doi:10.1080/01621459.1952.10483441)
69. Rosner B. 2011 Multisample inference. In *Fundamentals of biostatistics* (ed. M Taylor), pp. 516–576, 7th edn. Boston, MA: Brooks/Cole.
70. Dunn OJ. 1964 Multiple comparisons using rank sums. *Technometrics* **6**, 241–252. (doi:10.1080/00401706.1964.10490181)
71. Holm S. 1979 A simple sequentially rejective multiple test procedure. *Scandanavian J. Stat.* **6**, 65–70.
72. Charles H, Godfray J, Garnett T. 2014 Food security and sustainable intensification. *Phil. Trans. R. Soc. B* **369**, 1639.
73. Eckert S, Kohler S. 2014 Urbanization and health in developing countries: a systematic review. *World Health Popul.* **15**, 7–20. (doi:10.12927/whp.2014.23722)
74. Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW. 2004 The global distribution and population risk of malaria: past, present, and future. *Lancet Infect. Dis.* **4**, 327–326. (doi:10.1016/S1473-3099(04)01043-6)
75. Hanjra MA, Qureshi ME. 2010 Global water crisis and future food security in an era of climate change. *Food Policy* **35**, 365–377. (doi:10.1016/j.foodpol.2010.05.006)
76. Cohen B. 2006 Urbanization in developing countries: current trends, future projections, and key challenges for sustainability. *Technol. Soc.* **28**, 63–80. (doi:10.1016/j.techsoc.2005.10.005)
77. Lopez E, Bocco G, Mendoza M, Duhau E. 2001 Predicting land-cover and land-use change in the urban fringe: a case in Morelia city, Mexico. *Landsc. Urban Plann.* **55**, 271–285. (doi:10.1016/S0169-2046(01)00160-8)
78. Chabaeva A, Civco DL, Prisloe S. 2004 Development of a population density regression model to calculate imperviousness. In *ASPRS Annual Conference Proceedings*, *Denver, CO, 23–28 May 2004*. Bethesda, MD: American Society for Photogrammetry & Remote Sensing (ASPRS).
79. Jat MK, Garg PK, Khare D. 2008 Monitoring and modelling of urban sprawl using remote sensing and GIS techniques. *Int. J. Appl. Earth Obs. Geoinf.* **10**, 26–43. (doi:10.1016/j.jag.2007.04.002)
80. Sante I, Garcia AM, Miranda D, Crecente R. 2010 Cellular automata models for the simulation of real-world urban processes: a review and analysis. *Landsc. Urban Plann.* **96**, 108–122. (doi:10.1016/j.landurbplan.2010.03.001)
81. Small C, Nicholls RJ. 2003 A global analysis of human settlement in coastal zones. *Coast Res.* **19**, 584–599.
82. Cohen JE, Small C, Vitousek PM, Mooney HA. 1997 Estimates of coastal populations. *Science* **278**, 1209–1213. (doi:10.1126/science.278.5341.1209c)
83. Ericson JP, Vorosmarty CJ, Dingman SL, Ward LG, Meybeck M. 2006 Effective sea-level rise and deltas: causes of change and human dimension implications. *Glob. Planet Change* **50**, 63–82. (doi:10.1016/j.gloplacha.2005.07.004)
84. Huang B, Xie C, Tay R. 2010 Support vector machines for urban growth modeling. *Geoinformatica* **14**, 83–99. (doi:10.1007/s10707-009-0077-4)
85. Thapa RB, Murayama Y. 2011 Urban growth modeling of Kathmandu metropolitan region, Nepal. *Comput. Environ. Urban Syst.* **35**, 25–34. (doi:10.1016/j.compenvurbsys.2010.07.005)
86. Kates RW. 1971 Natural hazard in human ecological perspective: hypotheses and models. *Econ. Geogr.* **47**, 438–451. (doi:10.2307/142820)
87. Neis P, Zipf A. 2012 Analyzing the contributor activity of a volunteered geographic information project—the case of OpenStreetMap. *ISPRS Int. J. Geo-Inf.* **1**, 146–165. (doi:10.3390/ijgi1020146)
88. Haklay M. 2010 How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environ. Plan B. Urban Anal. City Sci.* **37**, 682–703. (doi:10.1068/b35097)
89. Openshaw S. 1984 *The modifiable areal unit problem*. Concepts and Techniques in Modern Geography no. 38. Norwich, UK: Geo Books.
90. Carrão H, Naumann G, Barbosa P. 2016 Mapping global patterns of drought risk: an empirical framework based on sub-national estimates of hazard, exposure and vulnerability. *Glob. Environ. Change* **39**, 108–124. (doi:10.1016/j.gloenvcha.2016.04.012)
91. Gleeson T, Befus KM, Jasechko S, Luijendijk E, Cardenas MB. 2015 The global volume and distribution of modern groundwater. *Nat. Geosci.* **9**, 161–167. (doi:10.1038/ngeo2590)
92. Davis KF, Rulli MC, Seveso A, D'Odorico P. 2017 Increased food production and reduced water use through optimized crop distribution. *Nat. Geosci.* **10**, 919–924. (doi:10.1038/s41561-017-0004-5)
93. Hansen MCC *et al.* 2013 High-resolution global maps of 21st-century forest cover change. *Science* **342**, 850–854. (doi:10.1126/science.1244693)
94. Gaughan AE *et al.* 2016 Spatiotemporal patterns of population in mainland China, 1990 to 2010. *Sci. Data* **3**, 160005. (doi:10.1038/sdata.2016.5)