University of Southampton

Faculty of Environmental and Life Sciences

Geography and Environmental Sciences

Modelling Global Human Settlement to Better Inform Annual Population Modelling

Volume 1 of 1

by

Jeremiah Joseph Nieves

ORCID ID 0000-0002-7423-1341

Thesis for the degree of Doctor of Philosophy

July 2020

University of Southampton

Abstract

Faculty of Environmental and Life Sciences

Geography and Environmental Sciences

Thesis for the degree of Doctor of Philosophy

Modelling Global Human Settlement to Better Inform Annual Population Modelling

by

Jeremiah Joseph Nieves

Since 1950, the world's population has shifted from being largely rural to majority urbanised. This trend of increasing urbanisation of population and increasing land use transitions promoting the growth of settlements and the built-environment, are expected to continue in future decades, particularly in low- and middle-income countries. These trends are accompanied by rapidly shifting subnational demographics and spatial distributions of populations, even within urbanised areas. Accurate and timely data is required to develop adaptive strategies for these shifting trends and minimising potential negative impacts. While multi-temporal, high-resolution datasets of built-settlement extent have become globally available, there remain gaps in their coverage and globally consistent methods of predicting future built-settlement expansion at regular intervals have not kept pace with these new data.

This thesis develops and validates a country-specific yet globally applicable means of annually interpolating built-settlement extents and projecting built-settlement extents into the near future using relative changes in subnational population and lights at night radiance. Additionally, I demonstrate the utility of this modelling framework within a global population modelling context across a period of 13 years. This thesis improves upon previous urban growth modelling approaches by demonstrating that relative changes in population can be sufficient, in and of themselves and as causal proxies for changes in economics, for accurately predicting areas undergoing built-settlement expansion across time and space. Additionally, this thesis validates its predictions at the pixel level, something not done by previous global urban and settlement modelling approaches. By addressing the limits that exist within current global urban modelling approaches, such as large or specific data requirements and subjective assumptions of growth factors/parameters, the modelling frameworks presented in this thesis allows for more consistent, frequent, and accurate built-settlement predictions. By extension, these accurate, time-specific built-settlement predictions allow for better, time-specific population mapping across the globe. Improved knowing of where and when built-settlement appeared allows for further investigations into arable land use consumption in relation to population dynamics, temporally fine-scale changes in population distributions across space in relation to climate change stresses, built-settlement expansion and greenhouse gas emissions, and trends in built-settlement expansion in relation to sea level rise, to name a few.

Preface

I began working with WorldPop in 2014 as a Researcher while carrying out my Master's degree at the University of Louisville. During that time I had numerous discussions regarding the gaps in urban related data and their applications to population modelling with Dr. Forrest R. Stevens (University of Louisville, Kentucky, USA), Dr. Andrea E. Gaughan (University of Louisville, Kentucky, USA), and Dr. Catherine Linard (Université de Namur, Belgium). These discussions during my Master's led to my conceptualisation of this body of work and the modelling framework – the Built-Settlement Growth Model). The papers submitted as a part of this three-paper thesis are the results of this work and have either been published or are under review for publication. They are as follows.

1. **Nieves, J. J.**, Sorichetta, A., Linard, C., Bondarenko, M., Steele, J. E., Stevens, F. R., Gaughan, A. E., Carioli, A., Clarke, D. J., Esch, T., & A. J. Tatem. (2020). Annually modelling built-settlements between remotely-sensed observations using relative changes in subnational populations and lights at night. *Computers, Environment and Urban Systems*. 80. https://doi.org/10.1016/j.compenvurbsys.2019.101444.
2. **Nieves, J. J.**, Bondarenko, M. Sorichetta, A., Steele, J. E., Kerr, D., Carioli, A., Stevens, F. R., Gaughan, A. E., & A. J. Tatem. (2020). Annual projections of future built-settlement expansion using relative changes in projected small area population and short time-series of built-extents. *Remote Sensing* 12, 1545. doi:10.3390/rs12101545.
3. **Nieves, J. J.**, Bondarenko, M., Kerr, D., Ves, N., Yetman, G., Sinha, P., Clarke, D. J., Sorichetta, A., Stevens, F. R., Gaughan, A. E., & A. J. Tatem. (Under Review). Measuring the contribution of built-settlement data to global population mapping. *Social Sciences and Humanities Open*.

In 2016, WorldPop had the "Global High Resolution Population Denominators Project", more generally referred to as the "Global Project," to construct consistent and comparable global maps of population distributions and demographics for the 1990-2020 period at 3 arc second (~100m) resolution and make these data openly and freely available. Such time-specific and demographically detailed data produced within a consistent framework was needed to better address public health monitoring and interventions, amongst other secondary applications, and as such was funded by the Bill and Melinda Gates Foundation (OPP1134076) and the Institute for Health Metrics and Evaluation (IHME).

At the time of project initiation, observed environmental covariate data, necessary for the population modelling, was largely unavailable past 2014 and was largely unavailable at annual time points between 2000-2014. This left an unfilled gap for important predictors of population, i.e. annual urban extents, between 2000-2014 and prediction of those covariates' spatial distributions and values from 2015-2020. This project specific gap hinted at a larger temporal gap in urban data that could be used for better population modelling and a variety of other applications which may require spatio-temporally detailed urban extents that are globally comparable. And, so, I developed my thesis topic around addressing this gap.

Looking to increase the impact of my thesis work and leverage the large amount (>10TeraBytes) of geospatial covariates being produced under the Global Project's resources, I closely partnered with the project. In this capacity, I carried out the following:

- Contributed to designing the spatial data infrastructure and spatial toolkits
- Had input on what covariates to process and how to process them, specifically those regarding representation of urban and the built-environment
- Trained several postdocs and PhD students in disaggregative population modelling
- Co-coded the revised population modelling scripts with Dr. Maksym Bondarenko (University of Southampton, UK)
- Co-produced an R package, documentation, and publication with Dr. Maksym Bondarenko
- Provided the Global Project with an early version of my urban growth model and documentation on how to scale it for production
- Co-supervised and participated in the production of the modelled urban datasets and modelled population datasets with Dr. Alessandro Sorichetta (University of Southampton, UK) and Dr. Maksym Bondarenko (University of Southampton, UK)

However, the Built-Settlement Growth Model framework, the focus of this thesis, was my conceptualisation and programming work.

Papers 1, 2, & 3 were undertaken under a ESRC Fellowship, for which I was Principal Investigator. I developed the initial concept, managed the programme of research, provided the intellectual direction of analyses, carried out the analyses, and led the preparation and submission of manuscripts. Dr. Andrea E. Gaughan, Dr. Forrest R. Stevens, and Dr. Catherine Linard were listed as co-authors throughout due to the large influence our early conversations (2014-2016) had on my final conceptualisation of the Built-Settlement Growth Model presented in this work.

Given the close work with the Global Project, many programmatic tools and frameworks for interfacing with the common geospatial data repository were used within my Built-Settlement Growth Model scripts. Additionally, since early versions of my model were utilised in the Global Project, several special cases and errors were found and corrected during the production process. For these reasons, Dr. Maksym Bondarenko, who I worked with side-by-side on Global Project, is listed as co-author on papers 1, 2 & 3.

Given the interdisciplinary nature of my work, guidance on utilising splines and growth curves was provided to me by Alessandra Carioli (University of Southampton, UK); therefore, she is listed as co-author on paper 1 & 2. Dr. Thomas Esch (DLR, Munich, Germany) provided both data and insight into state-of-the-art remote sensing datasets and therefore he is listed as co-author on paper 1. In paper 2, David Kerr (University of Southampton, UK) provided a programmatic tool for efficiently extracting built-settlement population for use in my analyses and carried out the extraction for me.

Given that paper 3 is a meta-analysis of population models for the globe from 2000-2020, many people were involved in the production of those models. Greg Yetman (CIESIN, University of Columbia, New York, USA), Dr. Parmanand Sinha (University of Louisville, Kentucky, USA), Nikolaos Ves (University of Southampton, UK), and David Kerr produced many of the countries' population and modelled urban datasets, using my Built-Settlement Growth Model, under the guidance and supervision of myself, Dr. Maksym Bondarenko, and Dr. Alessandro Sorichetta. Both Dr. Maksym Bondarenko and I also produced these datasets. They are all listed as co-authors for this reason.

Dr. Jessica E. Steele (University of Southampton, UK) provided supervision for papers 1 & 2. Dr. Donna J. Clarke (University of Southampton, UK), Dr.

Alessandro Sorichetta, and Dr. Andrew J. Tatem (University of Southampton, UK) provided supervision for papers 1, 2, & 3. All first drafts of papers were written by myself with comments and suggestions provided by co-authors and supervisors.

Additional papers under the Global Project where I am a co-author are either in preparation or are published as below.

Sinha, P. Gaughan, A. E., Stevens, F. R., **Nieves, J. J.**, Sorichetta, A. & A. J. Tatem. (2019). Assessing the spatial sensitivity of a random forest model: Application in gridded population modelling. *Computers, Environment and Urban Systems* 75: 123-145. doi: 10.1016/j.compenvurbsys.2019.01.006

Lloyd, C. T., Chamberlain, H. Kerr, D., Yetman, G. Pistolesi, L., Stevens, F. R., Gaughan, A. E., **Nieves, J. J.,** Hornby, G. MacManus, K., Sinha, P., Bondarenko, M., Sorichetta, A., & A. J. Tatem. (2019). Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. *Big Earth Data* 3(2). doi: 10.1080/20964471.2019.1625151

Bondarenko, M., **Nieves J. J. (co-primary author)**, Sorichetta, A., Stevens, F. R., Gaughan, A. E., & A. J. Tatem. (In preparation). wpRF: A package for random forest-informed population mapping. *Journal of Statistical Software.*

Table of Contents

Table of Tables

Table of Figures

List of Accompanying Materials

“BSGMiv1a Research Data and Code”
https://data.mendeley.com/datasets/f366zsg6hh/3
DOI: 10.17632/f366zsg6hh.3

“BSGMe Research Data and Code”
https://data.mendeley.com/datasets/cm6bnzvzfj/1
DOI: 10.17632/cm6bnzvzfj.1

“Measuring the contribution of built-settlement data to global population mapping - Supplementary Materials”
https://data.mendeley.com/datasets/2pxhmnxmnb/1
DOI: 10.17632/2pxhmnxmnb.1

Academic Thesis: Declaration of Authorship

I, Jeremiah Joseph Nieves ....

declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

Modelling Global Built-Settlement for Improving Time-Specific Disaggregative Population Modelling ....

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have consulted the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. Parts of this work have been published as:
   Nieves, J. J., Sorichetta, A., Linard, C., Bondarenko, M., Steele, J. E., Stevens, F. R., Gaughan, A. E., Carioli, A., Clarke, D. J., Esch, T., & A. J. Tatem. 2020. “Annually modelling built-settlements between remotely-sensed observations using relative changes in subnational populations and lights at night”. *Computers, Environment and Urban Systems*, 80.
   DOI: 10.1016/j.compenvurbsys.2019.101444

   Nieves, J. J., Bondarenko, M., Sorichetta, A., Steele, J. E., Kerr, D., Carioli, A., Stevens, F. R., Gaughan, A. E., & A. J. Tatem. 2020. “Predicting near-future built-settlement expansion using relative changes in small area populations”. *Remote Sensing*, 12(10):1545.
   DOI: 10.3390/rs12101545

Signed: ……………………………………………………………….

Date: ………………………………………………………………………………………………

Acknowledgements

I first need to thank my father, Jeremiah Nieves, Jr. He grew up impoverished and left Puerto Rico at 19 to join the military, helping financially support his siblings' pursuit of higher education at the expense of him foregoing his own. He later suspended his studies again to better support my sister's and mine education and to provide for us in ways he had not been, working two-three jobs to this day. He has always emphasised the value of education and the importance of using knowledge and talents to better understand and help others. While I have become cynical of many things, he has and will always remain a hero in my eyes. The things he has managed to accomplish, and his selflessness cannot be adequately summarised by anything less. I won't risk underestimating what else he will accomplish in his life, but if nothing else I wish him to know that this Doctorate is just as much his accomplishment as it is mine and I will be reminded of that every time I see "Dr. Jeremiah Nieves".

I need to thank my mother, Evelyn Thornberry Joly, who was always there to support me when the going got tough. We haven't always agreed or even understood each other, but you always served as an example of unconditional love, acceptance, and support. I also want to acknowledge the love and support of my grandparents Rachel Ann Thornberry and Joseph Lawrence "Larry" Thornberry ("Magar" and "Pagar"). I'll always cherish the times spent bird watching and drinking coffee on the front porch with Magar. And I have Pagar to thank for my love of educating; I didn't realise it at the time, but his example of patience and understanding while having me watch and learn in the barn and workshop have carried through to how I love to help and teach others.

For Nick and Cori Ruktanonchai, I still owe so much. You made crossing an ocean with two checked bags much less daunting than it could have been and provided a bit of Appalachia away from home. May shed doors never impede your life goals again.

I am also indebted to Maksym Bondarenko, Alessandra Carioli, and Kristine Nilsen who always made time to teach me new and valuable techniques, to mentor me as a researcher, and to promote my work and self-worth. You all really defined my PhD experience and I will forever be grateful.

Acknowledgements

I would be remiss if I didn’t take time to appreciate Donna J. Clarke, who selflessly mentored and supervised me unofficially for years before officially supervising me. In addition to this, you opened your home and welcomed me in as family, for which I will forever be grateful. While many other things have changed over these four years, I could always know that there was at least one person in the UK who’d pick up a shovel to lend a hand.

I also need to thank Forrest R. Stevens, Andrea E. Gaughan, and Catherine Linard who have continued to mentor me from afar and encouraged me to pursue a PhD. The seed of our discussions many years ago, on populations and urban environments, has finally sprouted and I look forward to seeing it continue to grow. Similarly, I wish to thank Thomas Esch who has generously and always made himself available to share perspectives on urban environment datasets and future directions of the field.

Lastly, I need to thank Andrew J. Tatem, Alessandro Sorichetta, and Jessica E. Steele for their supervision. Ale, even though you were overtasked you always found time to give detailed feedback. Andy, without your generosity and assistance in finding resources, my work would have had far less meaningful impact. Jess, I learned much from you on navigating and managing people and projects, which I am certain will serve me very well in the future. May we never enter the “Cave du Roy”.

I’m sure there are many others I could continue to thank for many more pages, and it pains me to not continue to write. I wish those who have been a part of this journey with me, no matter how long or brief, that I immensely appreciate them and, whether they know it or not, they impacted my life in meaningful ways. I love all y’all.

Definitions and Abbreviations

ARIMA – Auto-Regressive Integrate Moving Average

ARMA – Auto-Regressive Moving Average

BS – Built-Settlement

BSGMi – Built-Settlement Growth Model, interpolative

BSGMe – Built-Settlement Growth Model, extrapolative

ESA – European Space Agency

ESA CCI – European Space Agency Climate Change Initiative

ETS – Error Trends and Seasonality

GHSL – Global Human Settlement Layer

GIS – Geographic Information System

GLM – Generalised Linear Model

GUF – Global Urban Footprint

HBASE – Human Built-up And Settlement Extent

HRSL – High Resolution Settlement Layer

JRC – Joint Research Centre

OBIA – Object-Based Image Analysis

OSM – OpenStreetMap

PER.INC.MSE – Percent Increase in the Mean Square Error

RF – Random Forest

RS – Remote Sensing

SAR – Synthetic Aperture Radar

VGI – Volunteered Geographic Information

Chapter 1 Introduction to the Published Works

1.1 Introduction

While the global population growth rate has slowed to near 1950 levels, the total population is expected to increase to 9.7 billion by 2050 (United Nations, 2018, 2019). Over half of this growth is anticipated to occur in sub-Saharan Africa and within 47 of the lowest-income countries (United Nations, 2018, 2019). Combined with this are changing demographic compositions, with all countries experiencing aging populations. Higher-income countries currently have larger proportions of their population over 65 years of age (United Nations, 2018, 2019). This has implications for dependency ratios, i.e. the number of "working age persons" (15-64 years) per person under 15 and 65 years and older. This impacts the distribution and characteristics of labour and economics, which feeds into more potential migration flows (Montgomery *et al.*, 2003; Pezzulo *et al.*, 2017; United Nations, 2018, 2019).

In 2018, 55 percent of the world's population lived in urbanised areas, but this is projected to increase to 68 percent in 2050 (United Nations, 2019). This growth is primarily due to natural population growth, continued rural to urban migration, and the conversion of rural to urban land (Ledent, 1982; Angel *et al.*, 2011; United Nations, 2019). Similar to overall population trends, the majority of the urban growth anticipated to occur by 2050 will be in low- and middle-income countries, with 90 percent of the projected growth to occur in Asia and Africa (United Nations, 2018).

As the magnitude, demographic composition, and spatial distributions of populations change, issues adequately addressing sustainable development and urbanisation are dependent upon better understanding past, current, and future urbanisation trends (van Vliet, Eitelberg and Verburg, 2017; Zoraghein and Leyk, 2019; Ehrlich, Balk and Sliuzas, 2020). The rate of growth and magnitude of urbanisation requires greater information about urban areas and settlement, including higher frequency observations of urban areas (Hoalst-Pullen and Patterson, 2011; Acuto, Parnell and Seto, 2018; Zhu *et al.*, 2019). However, “it remains very important to take a settlement-based approach in studying the processes and patterns of demographic, economic, and social development”

(Champion and Hugo, 2017, p. xxi), with many settlements contributing to larger urbanised areas. Having settlement data with high spatial resolution, high temporal frequency, and broad temporal coverage is necessary to better understand the spatial distribution of human activities, societal processes, and human environment interactions (Ehrlich, Balk and Sliuzas, 2020).Settlement data with high spatial resolution have only recently become available (Pesaresi *et al.*, 2013, 2016; Esch *et al.*, 2013; Corbane *et al.*, 2017; Florczyk *et al.*, 2019). However, their temporal frequency and coverage are still sparse (Florczyk *et al.*, 2019).

More timely predictions and greater understanding of settlement expansion and urbanisation trends are necessary to minimise potential adverse outcomes and environmental impacts, and maximising the benefits that can come from urbanisation (Stephenson, Newman and Mayhew, 2010; Hoalst-Pullen and Patterson, 2011; Sverdlik, 2011; K C Seto, Guneralp and Hutyra, 2012; Eckert and Kohler, 2014; United Nations, 2018; Zhu *et al.*, 2019). Having such data could allow for more sustainable urbanisation via monitoring the preservation of arable land (Schaldach *et al.*, 2011; van Asselen and Verburg, 2013; van Vliet, Eitelberg and Verburg, 2017), reduce healthcare inequalities in unplanned and “slum” settlements (Vlahov *et al.*, 2007; Ezeh *et al.*, 2017), and adapting climate change mitigation strategies in the face of rapidly urbanising landscapes and populations(McGranahan, Balk and Anderson, 2007; McDonald *et al.*, 2011).

Addressing the issues of settlement growth and urbanisation and parallel changes in populations and their distributions requires more frequent and accurate population mapping (Freire *et al.*, 2020). Better mapping of population requires more frequent, consistent, and comparable mapping of the human footprint on earth, i.e. settlements (Balk *et al.*, 2006; Hoalst-Pullen and Patterson, 2011; Freire *et al.*, 2015, 2016; Champion and Hugo, 2017; Nieves *et al.*, 2017; Reed *et al.*, 2018; Zhu *et al.*, 2019; Zoraghein and Leyk, 2019; Ehrlich, Balk and Sliuzas, 2020; Stevens *et al.*, 2020). To achieve this, some key questions need to be addressed. First, to better predict these trends in population distributions and urbanisation, where and when urban/settlement expansion occurred across the globe in the past 15 years must be answered in a consistent and comparable manner with high temporal frequency (Hoalst-Pullen and Patterson, 2011; Champion and Hugo, 2017; Acuto, Parnell and Seto, 2018; Zhu *et al.*, 2019; Zoraghein and Leyk, 2019). With that answered, and expanding the available time series of urban/settlement extents, where and when urban growth is likely to occur in future years, based upon past and current trajectories, can be approached (Champion and Hugo, 2017; Zoraghein and Leyk, 2019). Then, how better predictions of the urban landscape can contribute to better population mapping can be investigated (Balk *et al.*, 2006; Freire *et al.*, 2015; Champion and Hugo, 2017; Nieves *et al.*, 2017; Ehrlich, Balk and Sliuzas, 2020).

However, the baseline data of past settlement extents and population distributions required to approach answering these questions in a consistent and comparable way across the globe do not currently exist with sufficient spatial or temporal detail (Acuto, Parnell and Seto, 2018; Zhu *et al.*, 2019; Ehrlich, Balk and Sliuzas, 2020). Small cities and towns are expected to experience high rates of growth, have large differences between settlements of differing size, and are traditionally the least well represented in current remote sensing data (Cohen, 2006; Champion and Hugo, 2017; Farrell, 2017; United Nations, 2018; Zhu *et al.*, 2019; Ehrlich, Balk and Sliuzas, 2020). Urban modelling would naturally present a potential answer to supplement existing urban datasets (K. C. Seto, Guneralp and Hutyra, 2012; Linard, Tatem and Gilbert, 2013; Zoraghein and Leyk, 2019; Ehrlich, Balk and Sliuzas, 2020). However, previously constructed urban modelling frameworks have limits, such as lack of specificity at subnational levels, that preclude their subsequent use to inform population mapping globally at high spatial resolution, particularly in data sparse low- and middle-income contexts (Champion and Hugo, 2017).

Here, I will introduce a novel and flexible urban modelling framework capable of global application and with the intended, but not exclusive, end use of improving population modelling. The focus of this thesis is on addressing the aforementioned questions and demonstrating that modelled urban extents can be useful in applied population modelling. Within this introduction, I will focus on four key themes. By understanding "Urban and Population Dynamics" I can better choose from the "Digital Representations of Built-Settlement" that are best suited for modelling population. By utilising the recent advances in digital representation of built-settlement, examining the limits of previous global and continental "Urban Growth Modelling" frameworks, and understanding "Built-Settlement in the Context of Modelling Populations", I develop an urban/built-settlement modelling framework that is optimised for better modelling populations.

1.2 Theme I: Urban and Population Dynamics

1.2.1 Definitions of Urban

Urban has been defined in many ways across many fields with different definitions existing within the same field, depending upon the specific application. Many countries define urban as a function of some population magnitude/density threshold, administrative jurisdictions, or functional economic areas and activities (Montgomery *et al.*, 2003; United Nations, 2015, 2018). While not conducive to applications requiring global consistency in definitions (Potere and Schneider, 2007), none of these definitions of the concept of urban are objectively wrong. The formal, yet vague, definition of urban is simply "of, relating to, characteristic of, or constituting a city" (Merriam-Webster, 2019). And, typically, what is not considered urban is simply classified as rural (Montgomery *et al.*, 2003; Champion and Hugo, 2017). However, urban is a complex amalgamation of the physical environment, population, economics, movement and connectivity, and their interactions (Burgess, 1925; Hoyt, 1939; Harris and Ullman, 1945; Gottman, 1957; Von Thunen, 1966; Zelinsky, 1971; Ledent, 1982; Berechman and Gordon, 1986; Southworth, 1995; Pozzi and Small, 2005; Cohen, 2006; Haas, 2010; Schneider, Friedl and Potere, 2010; Angel *et al.*, 2011). As recently as 1970, the number of application-based definitions of urban, was as little as three (Champion and Hugo, 2017). While the demand for more fit-for-purpose definitions of urban has continually grown, the geographical frameworks for collecting data on these definitions are often not built accordingly, risking misrepresentation of what is occurring and limiting future predictions of urban and population growth (Champion and Hugo, 2017, p. xxiii).

Needing an operationalised, measurable, and consistent definition of urban, many studies have used a definition based solely upon physical features observable from remotely-sensed imagery (Schneider *et al.*, 2003; Pozzi and Small, 2005; Small, Pozzi and Elvidge, 2005; Cheriyadat *et al.*, 2007; Potere and Schneider, 2007; Potere *et al.*, 2009; Small, 2009, 2016; Florczyk *et al.*, 2019). This has produced data that is broadly encompassed by the term "built-environment" (Table 1). Remote sensing (RS) refers to the measurement of the Earth's surface using sensors (e.g. satellite or aircraft based) that measure electromagnetic radiation from a distance. While the concept of urban is rich, the definition of urban land from a RS perspective is typically either thematic based, focused on anthropogenic and impervious land cover, or object based, focused on the urban features that comprise the thematic urban space such as transportation networks and buildings. Balk et al. (2018) have shown that, in certain contexts, RS-based definitions of the physical aspect of urban have large overlap with the population and density-based facets and definitions of urban. Because of the complex intermarriage of contributions from various fields to the concept of urban, any discussion of urban within an interdisciplinary context can easily become confusing. For consistency and clarity, moving forward I adopt the words and definitions outlined in Table 1 and deviate only with explicit description.

Table 1 Terms and definitions as related to urban and built-environment.

Terms	Definition	Additional Notes
Built Environment	Broad concept consisting of anthropogenic land covers and features	Includes impervious surfaces, buildings, lights, earthen structures, etc.
Built-Settlement	"Enclosed constructions above ground which are intended or used 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" (Pesaresi *et al*., 2013)	As detected by remotely sensed imagery, typically limited to buildings
Land Cover	The physical surface of the earth as observed *in situ* or via remotely sensed observations	
Land Use	The anthropogenic activities associated with a given area of land	A single land cover could have many land uses. While land cover is easily detected by remotely sensed observations, land use is typically not detectable or classifiable, e.g. how to discern between an aquaculture pond and a natural pond.
Urban	Broad concept incorporating the physical environment, population, economics, and movement and connectivity	
Urban Feature	A subset of the built environment, but containing built-settlement, that focuses on anthropogenic objects found within built environments	As detected by remotely sensed imagery, could be anything from roads, buildings, other infrastructure objects, etc.
Urban Growth	The growth of population within urban areas and the population which find themselves being reclassified as urban due to urbanisation	Urban growth, a population focused concept, can occur without further urbanisation due to natural growth and in migration to urban areas
Urbanisation	The transitional process from an "agrarian to an industrial society" typically accompanied by changes in the distribution of economic activity, an expansion and or intensification of the built environment and, typically, a densification of the population distribution across space (Zelinsky, 1971; Ledent, 1982)	

1.2.2 The Urban, Population, and Economics Triad

Generally, in the 19th and 20th centuries surplus labour from rural areas migrated to larger settlements where this labour was applied to producing goods, modernising agriculture, and developing specialised services (Davis, 1965; Anas, Arnott and Small, 1998; Montgomery *et al.*, 2003; Farrell, 2017). Parallel advances in transportation lowered the cost of moving people, goods, and information further facilitated growth and made the services provided by cities, e.g. health care and treated water, accessible to a larger population (Preston and van de Walle, 1978; van Poppel and van der Heijden, 1997; Montgomery *et al.*, 2003)

This economic activity and advances in transportation were largely the originator of and sustained the growth of cities during this period, but in more contemporary times city and urban growth has become more detached from economic activity and, consequently, rural-urban migration and becoming more a consequence of natural population growth and increases in health sciences and services (Davis, 1965; Preston and van de Walle, 1978; Ledent, 1982; van Poppel and van der Heijden, 1997; Montgomery *et al.*, 2003; Dyson, 2011; Farrell, 2017). While these factors are described at the city scale, there is an interdependence upon inter-city economic activity as well as the influence of regional, national, and international policies (Benziger, 1996; Montgomery *et al.*, 2003; Joss, Cowley and Tomozeiu, 2013). Further, "urban areas" and "urban populations" are often growing simply due to the reclassification of non-urban areas to urban (Cohen, 2006; Balk *et al.*, 2018; Jones, Balk and Leyk, 2020). While urban growth, the absolute growth of population within urban areas, and urbanisation, the growth of the proportion of people living in urban areas, often occur simultaneously, these processes can occur without each other (Davis, 1965; Montgomery *et al.*, 2003; Farrell, 2017). Indeed, the footprint of cities and settlements can grow with little economic or demographic input, but rather be the result of an increased demand of services and living space (Montgomery *et al.*, 2003; Ehrlich, Balk and Sliuzas, 2020).

While cities can have benefits, they also can have costs such as higher exposure to pollution and crime, higher cost of property, higher commuting times, and more crowded living conditions (Preston and van de Walle, 1978; Glaeser, 1998). Some cities are actually experiencing decline across the economic, infrastructure, and population spectrums (Glaeser, 1998; Oswalt, Rieniets and Schirmel, 2006; Hollander and Németh, 2011). However, this decline, if underlying a process of deconcentration of a certain city's "primacy," can stimulate longer term economic growth (Henderson, 2002). But, if an area of a city becomes depopulated, the physical structures, such as buildings and roads, may remain for years, decades, or centuries depending upon the construction, environmental factors, and if there is planned demolition of such structures. This is to say that once an area has been converted to an urban or settled land cover, it is relatively inelastic to changing back to a more natural environment (Verburg *et al.*, 2002, 2004; Schaldach *et al.*, 2011; van Asselen and Verburg, 2013).

While the above factors are generally common across cities and urban areas as a whole, there drivers, patterns, and trajectories of cities and urban areas can vary quite widely across both space and time (Farrell, 2017). As shown, the factors that contribute to the growth of a city can also contribute to its decline or shrinkage and the way these factors may interact is highly context dependent and operates across a variety of spatial and temporal scales. Cities and urban areas do not exist in a vacuum; they are dependent, symbiotic, and in competition with rural areas with regards to land, natural resources, labour, capital, services, and infrastructure (Douglass, 1989; Kelly, 1998; Montgomery *et al.*, 2003). And as varied as city and urban growth was, and still is, there arose numerous theoretical models to try and understand these complex processes between urban areas, economics, and populations.

Many of the 19th and through the mid-20th century urban models were used as conceptual lenses through which urban development and the transition of urban form could be understood. In 1875, Von Thünen (1966) described land use through economics based upon land rent and transportation cost from a central market. Burgess (1925) expanded upon this and put forth the concentric zone model consisting of a Central Business District (CBD) surrounded by five concentric zones of differing utilitarian purposes: transition zone, inner suburbs, outer suburbs, and a commuter zone. Hoyt (1939), remaining with the idea of a CBD, proposed a series of wedges or sectors, as opposed to concentric rings, radiating from the CBD along lines of transportation/communication with specialised uses similar to Burgess's model. Harris and Ullman (1945) later moved away from the pivotal CBD by describing an urban area as a series of nuclei, that develop around an existing CBD, and around which specialised activities are focused. This has since developed into a broader more pluralistic concept of "polycentric cities" where spatially clustered, historically independent cities are physically connected via infrastructure and have a network of flows of people, goods, and ideas across various scales (Gottman, 1957; Kloosterman and Musterd, 2001; Parr, 2004; Green, 2007).

Regardless of their form, these conceptual models had the common elements of explaining the relationships between people, the built-environment, and economic activity and across space (Figure 1).

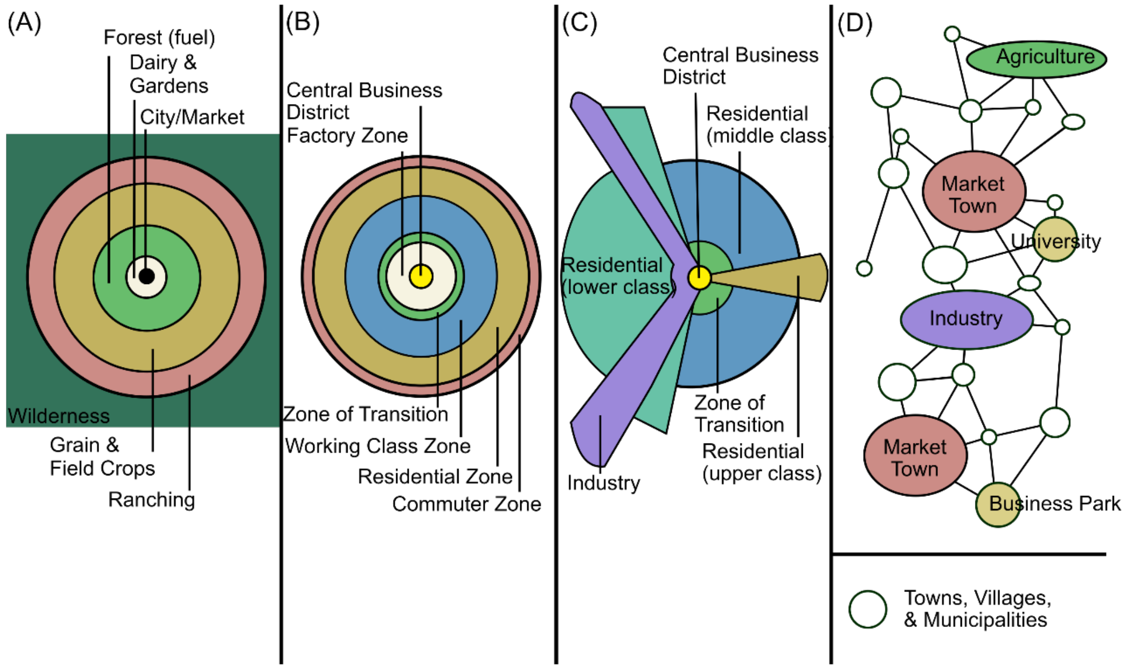


Figure 1 Generalised diagrams of the prominent 19th and 20th century urban conceptual models with (A) Von Thünen's Agricultural Land Use Model, (B) Burgess's Concentric Zone Model, (C) Hoyt's Sector Model, and (D) the Polycentric Model.

Despite the utility of these conceptual models, "...city and regional systems are entities far too complex to be understood through theory alone" and "...even the evolution of a single city presents issues of path dependence, historical lock-in, and agglomeration dynamics that render theoretical conclusions indeterminate. The difficulties involved in explaining a single city's growth trajectory are magnified many times when city systems and regions are considered" (Montgomery *et al.*, 2003, p. 58).

Given that there are established, although complex, causal linkages between population, the built-environment, and economic activity it would follow that these are factors I should consider in modelling any aspect of urban. This presents an issue when attempting to globally model as high-resolution, spatially or temporally, data on economic activity are not globally available below the national level (Gao and O'Neill, 2019, 2020).Logically, that leaves using available population data to predict the built-environment. This is because population is not only a driver of land cover changing from non-built to built (Ledent, 1982; Montgomery *et al.*, 2003; Cohen, 2006; Dyson, 2011), but, lacking better information, could be sufficient to predict changes in the built-environment and or potentially serve as a proxy for any economic drivers of these transitions. Given that built-environment data is consistently important in predicting populations across the globe (Nieves *et al.*, 2017), I hypothesise that the inverse can be true as well. In this work, I leverage the correlations, in data, that result from these complex and endogenous causal relationships. However, inferring the specific forms of these causal relationships is beyond the scope of this thesis.

Prior to utilising these relationships, I must further define what "urban" I am investigating. Even restricting the definition of urban to the built-environment as measured from RS-based imagery, there is still plurality in the definition. I first need to answer how I intend to classify and measure urban consistently across the globe.

1.3 Theme II: Digital Representations of Built-Settlement

Even by reducing the definitional scope of urban, the form of the built-environment can widely vary across space and time due to materials used, differences in urban morphology, and the surrounding environmental context (A Schneider and Woodcock, 2008; Small, 2009; Schneider, Friedl and Potere, 2010; Jilge *et al.*, 2019). The built-environment broadly encompasses the anthropogenic physical environment (Table 1) including urban features such as buildings, roads, runways, other impervious surfaces, and, depending on the spatial scale of the dataset, can include semi-natural/managed land covers such as golf courses or suburban lawns and gardens.

Currently, almost all digital past and present built-environment extent data is based upon RS imagery (e.g. satellite- or airplane-based imagery) or Volunteered Geographic Information (VGI) (e.g. OpenStreetMap), in the form of manually delineated areas of RS imagery. Initially, these datasets took the form of thematic classifications of urban land cover created from RS optical imagery, with the "urban" class typically capturing the "built-environment" (Table1). Such a class would incorporate urban features such as buildings, roads, runways, other anthropogenic impervious surfaces, and, sometimes erroneously, bare soil (Vogelmann *et al.*, 2001; Schneider *et al.*, 2003; Yang *et al.*, 2003; Bartholomé and Belward, 2005; Potere *et al.*, 2009; Schneider, Friedl and Potere, 2010; European Space Agency, 2013; UCL Geomatics, 2017). These datasets are typically at 30m or coarser in resolution. Later improvement used supporting information about the surrounding environment and vegetation during post-processing to help discern the true built-environment from the surrounding "noise" producing a global dataset at 500m resolution (Schneider, Friedl and Potere, 2010). Other recent development in thematic urban land cover datasets include a global impervious surface dataset at 30m resolution (Brown de Colstoun *et al.*, 2017b).

Coinciding with advances and availability in imagery, statistical methods, and computational resources, along with increases in layperson Geographic Information System (GIS) literacy, manually delineated urban feature data have become available and prevalent. These datasets include individual objects of the built-environment, such as building footprints and roads. Generally, manual delineation is too time consuming and expensive for producing urban extents at a global scale, much less for multiple time points. Most large delineation efforts, such as those undertaken by Humanitarian OpenStreetMap Team (HOT), are triggered by specific events, e.g. earthquakes or hurricanes (hotosm.org). Nevertheless, OpenStreetMap (OSM) is a large, popular, and important freely available repository of VGI containing manually defined vector-based urban feature data. However, the completeness, accuracy, and representation of the data in time is either unknown or variable (Haklay, 2010; Neis and Zipf, 2012; Linard *et al.*, 2014). In addition to OSM, many companies, governmental organisations, and others have produced similar datasets, although their availability is not always free and open. The advent of these manually delineated urban feature datasets provided a rich resource of training data for feature extraction algorithms and initiated a series of advances in the production of urban feature datasets.

While not exhaustive, I briefly cover notable recent advances in refining the extraction of urban features with global extent and across several time points.

- The European Commission Joint Research Centre (JRC) produced the 38m resolution Global Human Settlement Layer (GHSL) (Pesaresi *et al.*, 2013, 2016). This data set utilised a corner detection algorithm with symbolic learning methods to extract human-habitable and related structures from the entire Landsat catalogue of imagery, producing built-settlement extents for 1975, 1990, 2000, and 2014 at 38m resolution (Table 2) (Pesaresi *et al.*, 2013, 2016).
- The Global Urban Footprint (GUF) dataset utilised Synthetic Aperture Radar (SAR) imagery, along with post-processing based on OSM data, to globally extract vertical structures at 12.5m resolution, representing circa 2012 (Table 2) (Esch *et al.*, 2013).
- The European Space Agency's Climate Change Initiative (ESA CCI) land cover dataset has annual coverage at 300m from 1992-2019 (UCL Geomatics, 2017). It is unique in that it blends traditional multi-spectral thematic definitions of the built-environment, albeit only capturing for land cover change lasting at least two years, with the settlement extents of the GHSL and GUF datasets (Table 2) (UCL Geomatics, 2017).
- Facebook used a combination of manually delineated urban feature data and automated machine learning methods to extract urban features and produce the High Resolution Settlement Layer (HRSL). The HRSL dataset covers 33 countries, has 30m resolution, and represents the year 2015 (Table 2) (Facebook Connectivity Lab and Columbia University Center for International Earth Science Information Network - CIESIN, 2016).
- The Center for International Earth Science Information Network (CIESIN) used a method similar to Facebook to extract urban features globally from Landsat data, representing 2010, at 30m resolution, and post-processed them using OSM data to create the Human Built-up and Settlement Extent (HBASE) dataset (Table 2) (Brown de Colstoun *et al.*, 2017a).
- While not global in extent, Microsoft also used an automated convolutional neural network process to produce publicly available vector-based building footprints across the entire United States and added them to the OSM database (Table 2) (Microsoft, 2018).
- For 40 cities in Africa, Forget, Linard, and Gilbert (Forget, Linard and Gilbert, 2018) leveraged the information in OSM building footprint data to train a machine learning model with object based image analysis (OBIA) approaches (Table 2). This was used to extract urban features from a

combination of optical and SAR imagery at 12.5m for a series of time points between 1995 and 2015.

These latter two datasets are given as further examples of the strength of OBIA methods in urban feature extraction used in conjunction with VGI, such as OSM, and possible hints of future trends in urban feature extraction methods. These datasets and their characteristics are detailed in Table 2.

Table 2 Non-exhaustive list of RS thematic and feature of built-environment and built-settlement datasets with their characteristics.

Dataset	General Derivation Method	Urban Representation Type [a]	Sensor Type & Data	Spatial Extent	Spatial Resolution (meters)[c]	Temporal Resolution	Temporal Extent
ESA CCI Landcover	Thematic Image Classification	BE-BS	MS	Global	300	Annual	1992-2018
GHSL	Symbolic Image Learning	BS	MS	Global	38	Cross-sectional	1975, 1990, 2000, 2014
GUF	Support Vector and Heuristic Post-processing	UF-BS	MS & VGI	Global	12.5	Single Time Point	Circa 2012
HBASE	Convolutional Neural Network	BE-BS	MS	Global		Single Time Point	Circa 2015
HRSL	Convolutional Neural Network	BE-BS	MS			Single Time Point	Circa 2015
MAUPP	OBIA	BE-BS	MS & SAR	40 African Cities	12.5	Every 5 years	1995-2015
MODIS 500	Thematic Image Classification	BE	MS	Global	500	Single Time Point	2000
Microsoft Building Footprints	Artificial Neural Network Feature Extraction	UF-BS		3 countries	Vector	Single Time Point	Circa 2015-2018

[a] BE: Built-Environment BS: Built-Settlement UF: Urban Feature(s)
[b] MS: Multi-spectral SAR: Synthetic Aperture Radar VGI: Volunteered Geographic Information
[c] Approximate at the Equator

The primary trend in these advances is a shift towards higher spatial resolutions, with greater temporal frequency of the derived datasets (Florczyk *et al.*, 2019). Additionally, a conceptual shift occurred, from capturing the wider construct of the built-environment to capturing individual urban features or urban features that contribute to the concept of built-settlement (Florczyk *et al.*, 2019) (Table 2). This is particularly important for applications relating to human presence and activities as it is generally considered that a settlement based approach is the most appropriate for investigating demographic, economic, and societal processes (Champion and Hugo, 2017, p. xxi; Ehrlich, Balk and Sliuzas, 2020).

Meyer and Turner (1992, p. 47) give "settlement" to have one of the two meanings, depending on the overall conceptual lens adopted:

> "The category of settlement as a land use includes areas devoted to human habitation, transportation, and industry. As land cover, it incorporates highly altered surfaces such as buildings and pavement, but such cover represents only a portion of the total area that a land-use classification might accord to settlement."

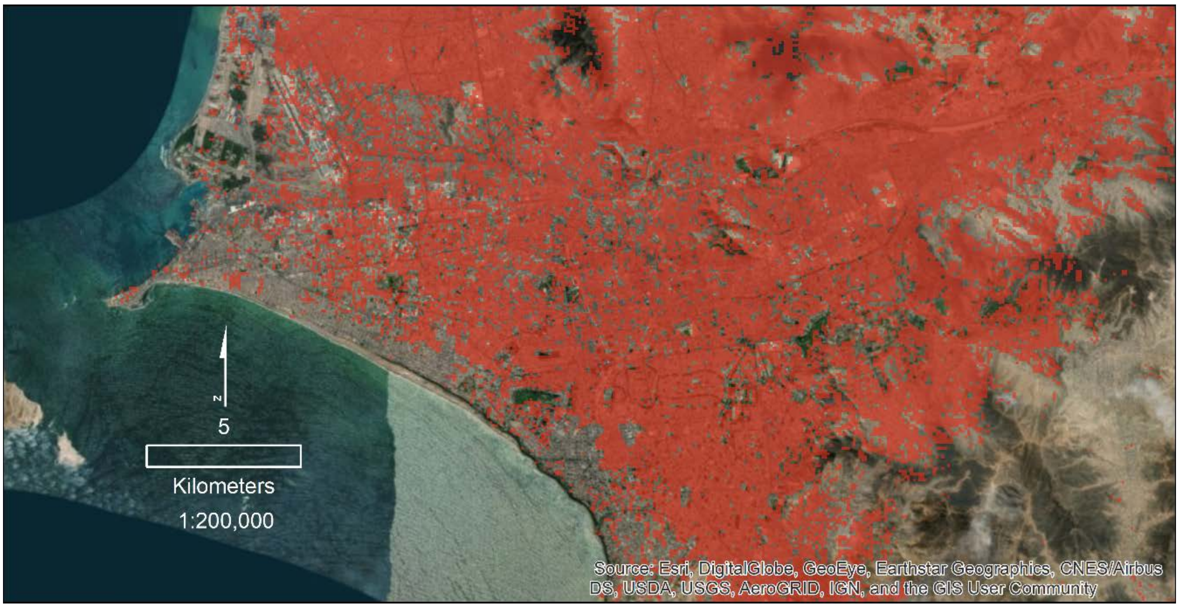
However, the physical realisation of settlement, i.e. the morphology, varies across space and time (Schneider *et al.*, 2003; A. Schneider and Woodcock, 2008; Small, 2009; Jilge *et al.*, 2019). These variabilities are a result of the land use, materials available, and the underlying urban processes that occur across numerous spatial and temporal scales (Zelinsky, 1971; Berechman and Gordon, 1986; Montgomery *et al.*, 2003; Dyson, 2011; Farrell, 2017; Small *et al.*, 2018). This variability is particularly apparent in RS data both between and within cities, as well as across regional and international scales (Small, 2009; Jilge *et al.*, 2019). See Figure 2 of Jilge *et al.* (2019) for an example of the coexistence of 24 different material classes, as detected by RS imagery, in a 2.4 km^2 area. Additionally, including "highly altered" surfaces such as pavement, roads, alongside buildings in my adopted definition of settlements would not be useful for my stated purpose of expanding the data availability of where humans are typically located, i.e. population mapping (Pesaresi *et al.*, 2013; Esch *et al.*, 2013; Freire *et al.*, 2015; Ehrlich, Balk and Sliuzas, 2020). Because of these two considerations, a further narrowing of the fit-for-purpose definition of settlement and urban features (in this case, buildings) I adopt is necessary.

Here, I adopt the definition of “built-settlement.” Built-settlement (BS) is based upon the concept of “built-up structure” in the Global Human Settlement Layer (GHSL) urban feature dataset (Table 2) and is defined as “enclosed constructions above ground which are intended or used 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” (Pesaresi *et al.*, 2013, p. 2108). This is similar to the concept of “built-up land” in Balk, Leyk, *et al.* (2018). More broadly, the concept of BS is an intersection between the urban features of the built-environment and population. Hereafter, I generalise BS to include other urban feature datasets that attempt to refine their thematic or feature set to better represent buildings associated with human activities and exclude more general impervious surfaces, such as roads, parking lots, and runways.

Indeed, remotely sensed derived datasets have recently shifted away from the field-based concept of urban land cover towards the object-based concept of urban features. Determining land use from remotely sensed imagery is often impossible. For instance, it may be difficult to differentiate from a commercial use building and a residential use building. Nevertheless, there are aspects of the urban physical environment that can be logically inferred to allow for specific land use such as human habitation. As seen in the above datasets, detection of right-angled corners is often associated with human-constructed buildings (Pesaresi *et al.*, 2013; Forget, Linard and Gilbert, 2018). Additionally, backscattering from SAR data allows for the detection of structures that are perpendicular to the Earth’s surface, that is, vertical constructions such as buildings in the GUF data set (Esch *et al.*, 2013). If the interest is in human settlements and human populations, then it is logical to adopt datasets that align with the definition of BS (Table 1).

BS datasets available with global extent are relatively new, with early versions of GHSL being the first, circa 2014, however there are still limits to these datasets, their input imagery, and associated methods. While these datasets can be accurate and useful, in that they are pre-processed to facilitate use by end users, the resource cost of production is quite large. For instance, GUF utilised SAR imagery captured between 2011-2013 at 12.5m resolution and handled over 400 TB of data to produce, in 2015, the single GUF time point representing 2012. This is a two-year time lag from having imagery to producing the final product involving extensive super-computing facilities (Esch *et al.*, 2018a). As implied, GUF represents only a single time point, circa 2012, and GHSL has no information between the four produced time points of 1975, 1990, 2000, and 2014. Even the follow-up dataset to GUF, the World Settlement Footprint (WSF) is predicted to only have observations every five years (Esch *et al.*, 2018a) and has been delayed by two years due to computational burdens and ensuring a sufficiently high level of quality outputs (Esch, 2019). This means there are gaps in the current temporal coverage of BS datasets, which precludes answering some of the larger questions I listed in the introduction.

There can also be gaps in the spatial coverage due to the characteristics of the RS sensors utilised as inputs for these datasets. For instance, cloud cover, suboptimal atmospheric conditions, or sensor errors, can result in missing areas within the resulting BS dataset. For example, in Lima, Peru (Figure 2) GHSL failed to detect an old area of the city at the 1975 time point and subsequently did not classify it as BS in later years.



Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AeroGRID, IGN, and the GIS User Community

Figure 2 Example of GHSL built-settlement extent data (red) in Lima, Peru where a part of the city pre-existing 1975 was not captured (south west coast). The artefact of the image tile extents where there were issues (e.g. atmospheric, sensor, etc.) is also apparent.

Compounding potential spatial gaps, there is generally an inverse relationship between the temporal resolution and the spatial resolution of global BS datasets. That is, the higher the temporal frequency of the BS dataset, the lower the spatial resolution is.

The ESA CCI dataset, which is BS-like, has an annual temporal resolution, but has a spatial resolution of 300m at the Equator (Table 2). The practical trade-off of this is that small settlements, or small area changes of larger settlements across time, are not detected by the ESA CCI dataset and there is a higher confusion of bare-surfaces with BS (UCL Geomatics, 2017). Which, if small to medium-sized settlements truly are experiencing the largest amounts of growth, this has significant implications of bias when using such a coarse spatial resolution. That is not to say a finer spatial resolution dataset like GHSL, which often identifies areas of turbid white water and dark objects against homogenous and relatively bright backgrounds as BS, or GUF, which erroneously identifies bridges and stacks of shipping containers as BS, are not without their own classification faults.

Regardless of their methods or input imagery, RS-based BS datasets promise to continue to be a rich source of information on urban systems in the future as imagery, methods, and computation resources improve and become more accessible. Even with globally consistent RS-based datasets, there are notable limits with RS-based BS data currently including:

i) Based upon the aforementioned BS data, an inverse relationship between spatial and temporal resolutions of the datasets

ii) Gaps in data due to hardware errors or atmospheric conditions producing unusable imagery (e.g. Landsat 7 scan line error or) (Figure 2) (Leyk *et al.*, 2019)

iii) Various spectral definitions of what constitutes the urban environment in addition to spectral signatures of the urban environment that varies across space (Potere and Schneider, 2007; Potere *et al.*, 2009; Small, 2009; Florczyk *et al.*, 2019; Jilge *et al.*, 2019)

iv) Can be computationally expensive to extract the urban extent features and typically requires training data that can be prohibitive to obtain on a global scale (Pesaresi *et al.*, 2013; ESA CCI, 2017; Forget, Linard and Gilbert, 2018; Esch *et al.*, 2018a)

v) Most urban growth occurs at smaller spatial scales, meaning that the currently available high-temporal resolution (i.e. low spatial resolution) data is not suitable nor able to capture the urban environmental changes, or existence of, smaller settlements and changes (Cohen, 2006; Champion and Hugo, 2017; Farrell, 2017; United Nations, 2018; Zhu *et al.*, 2019; Ehrlich, Balk and Sliuzas, 2020)

Additionally, by their observational nature, satellite-based data will never be able to see future, or past unobserved, extents. If the existing temporal gaps in these datasets are to be filled in and subsequently extended to project future BS extents, it would need to be without relatively large computational burdens and imagery requirements. This is where an urban growth/expansion model would appear to be the most tractable solution.

1.4 Theme III: Urban Growth Modelling

Within the conceptual urban models of the 19th and 20th century, the common elements of people, resources/economics, transportation, and space (Figure 2) were largely treated as static and revolved around aggregate levels of macro-economic "…relationships between various types of production and consumption…" (Batty, 2009, p. 52). However by the 1950s, with the advent of digital computers, urban models shifted from conceptual to computational in nature and have since spanned from macro- to micro-economic in their focus and from static to dynamic in their nature (Batty and Xie, 1994; Batty, 2008, 2009; Li and Gong, 2016a). This spectrum of urban modelling, ranging from top-down cross-sectional explanations of "city in equilibrium" to more bottom-up explanations of "urban location and behaviour" is displayed in Figure 3 (Batty, 2008, p. 3).

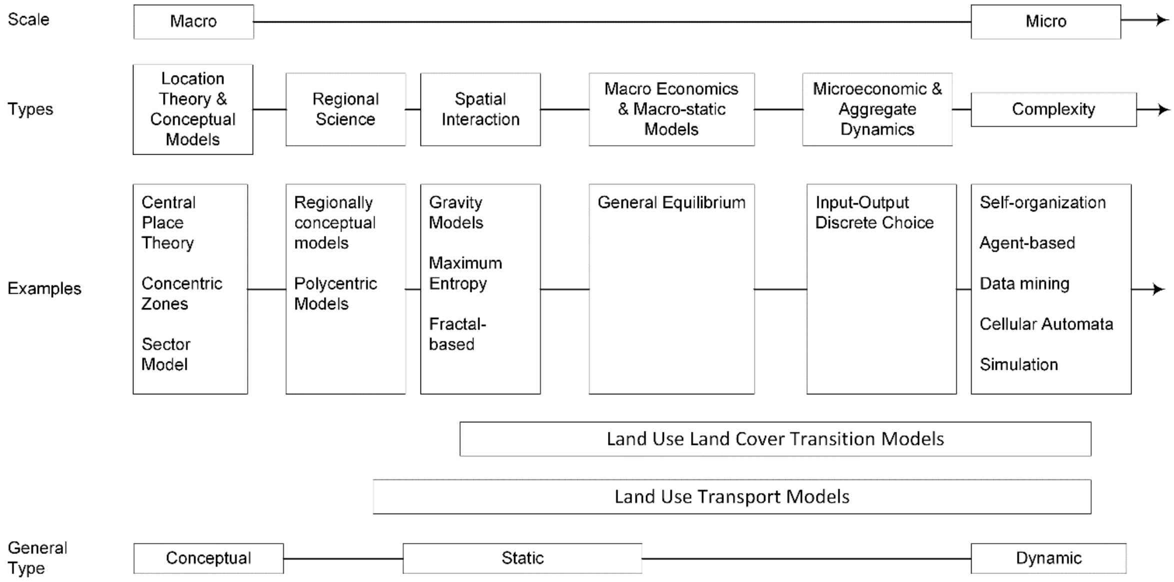


Figure 3 Continuous spectrum of urban models from top-down aggregate conceptual models (far left) to the computational top-down cross-sectional "city-in-equilibrium" models and the more bottom-up and dynamic models (far right) Model categories are broad and examples are non-exhaustive. Adapted from Li and Gong (2016a) Fig. 1 and Batty (2008) Fig. 1.

In the 1960s, a shift began towards modelling on a smaller scale and on city dynamics and growth, as opposed to a static cross-sectional representation of cities (Batty, 2008). In simpler terms, the focus shifted towards modelling the process of "urbanisation". Urbanisation, from a socio-economic view, is the transitional process "from an agrarian to an industrial society" (Ledent, 1982) with the process having three primary pathways: (1) natural growth of existing urban populations, (2) migration of population from rural to urban areas and, (3), the reclassification of rural areas to urban, due to changes in non-population aspects of the concept of urban such as spatial distribution of economic development and activities (Figure 1) (Zelinsky, 1971; Ledent, 1982). The process of urbanisation is not synonymous with population-centric concept of "urban growth", with urbanisation occurring contemporaneously with urban growth only if urban population grows more rapidly than rural population (Ledent, 1982).

With all the varying factors and drivers of urban, this eventually led to the three main cotemporary groups of urban growth models that incorporate both time and space: Land Use Transport (LUT), Agent-based, and Cellular Automata (CA) models (Batty, 2008; Li and Gong, 2016a).

LUT models focus on the zonal allocation of socio-economic activities, specifically around the idea that transportation networks influence location decisions of residential and economic land uses, ultimately leading to larger scale land use configurations and zoning (Berechman and Gordon, 1986; Southworth, 1995; Batty, 2009).

Agent-based models allow for individual human units or “agents” to have various decision making characteristics and interact with not only each other, but modelled aspects of the physical and socio-economic environment to make decisions on their spatial distribution (Sakoda, 1971; Schelling, 1971; Ferber, 1999; Benenson, 2004; Li and Gong, 2016a). These individual interactions and self-organising type behaviour and decisions give rise to the larger aggregate urban form and the more complex dynamics (Li and Gong, 2016a).

By the 1990s, cellular automata type models had come to the forefront of the urban modelling discipline (Batty, 2008, 2009). Because this thesis is global in extent and focuses on data that is globally available, I will focus on CA type models in more depth. CA models typically have less input data requirements in that they handle transportation in a much less nuanced manner, often do not directly handle economic activities, and do not generally attempt to, directly, model sociological factors.

1.4.1 Urban Cellular Automata Models

Cellular Automata (CA) models, first described by Alan Turing and later popularised by Von Neumann in the 1950s (Batty, 1997), are capable of modelling complex phenomena whose foundational components can be considered relatively simple (Wolfram, 1984; Batty, 1997). Tobler’s (1970) “Computer Movie” of Detroit can be considered the first application of CA to modelling urban environment growth. In the application of CA to modelling urban environment growth, the identical components of the CA typically take the form of the cells of a regular spatial grid, i.e. a raster (Sante *et al.*, 2010). Each cell has an initial value which changes based upon an applied set of deterministic rules and the variation in the pattern of resultant values depends upon the local conditions of neighbouring values and the values of data in and around a given cell (Wolfram, 1984; Sante *et al.*, 2010). These rules, which determine the transition probability of a cell, can be heuristic-based, dynamic or static across space, time, and scale, or determined by statistical means, such as machine learning methods (Clarke, Hoppen and Gaydos, 1997; White and Engelen, 1997, 2000; Verburg *et al.*, 2002; Sante *et al.*, 2010; Schaldach *et al.*, 2011; van Asselen and Verburg, 2013). These different types of transition rules were grouped into six general categories by Sante et al. (2010), which I describe in Table 3.

Table 3 General classes of transition rules per Sante et al. (2010)

Transition Rule Type	General Description
I	The state of a given cell is a function of the cell's current state and its neighbours' states
II	Assumes the primary driver of urban change can be captured as a "transition potential," (probability of a cell changing to a given land cover), which is a function of the cell's current state, its neighbours' states, land use constraining factors, and, sometimes, stochastic elements.
III	Rules based upon the "urban shape" having a basis in measures of landscape ecology and fractal theory
IV	Rules automatically constructed using machine learning methods and algorithms
V	Rules based on fuzzy-logic
VI	All other methods of generating rules including logical operations based upon assumptions, histogram distributions, and others.

Further, despite the variation in specificities, the influences that give rise to the transition rules are generalisable as compiled in Table 4 (Sante *et al.*, 2010).

Table 4 Common influences creating transition rules found in urban cellular-automata models (Sante *et al.*, 2010).

Influence Source	Description
Transition Rules	See Table 3
Model Objective	Is the model descriptive (capturing urban dynamics as related to factors), predictive (projecting future urban extent), or prescriptive (attempting to find optimal urban configuration)?
Cell Space	Various cell sizes can be utilised, but the cell size should be chosen based upon the size of the objects of interest in the urban landscape.
Cell States	Is the model simulating non-urban to urban transitions or accounting for urban transitions to and from multiple discrete non-urban classes?
Neighbourhood	Neighbourhood size and type will have an effect on the model output and should be taken into consideration with study area, cell space, and time frame.
Growth Constraint	How is urban growth generated? Endogenously or exogenously? Should there be rate restrictions?
Integration with other Models	Is the model incorporating outputs or informing other models, e.g. population models that then inform the urban growth magnitude or timing?
Calibration	Procedures to adjust transition rule parameters to produce the most accurate modelling of past urban transitions.
Validation	Methods of assessing modelled urban extent maps to corresponding real urban extent maps which can influence data availability for training of model.

The variations and specific transition rules of CA models for urban growth prediction are innumerable, see Sante et al. (2010) and Li and Gong (Li and Gong, 2016a) for a more detailed review.

CA urban growth prediction models typically focus on the physical environment, as opposed to the socio-economic environment, to determine the transition or non-transition of a cell from non-urban to urban, although most times the transportation aspect is often absent from such models (Batty, 2009). The gridded nature of the CA and its data inputs and outputs allows for ease of aggregating to any spatial unit for subsequent application and more importantly, for the modelling aspect, allows for efficient application parallel computing methods to large study areas (Sante *et al.*, 2010).

While urban CA models have greatly furthered the field of urban modelling and urban planning, there is a distinct regional bias to these studies, with the majority of them occurring in high- to middle-income countries (Seto *et al.*, 2011). Further, most models of spatial urban growth and transition have been city-specific or sub-national in scope, parameterisation, and data dependency (Clarke, Hoppen and Gaydos, 1997; White and Engelen, 1997; Clarke and Gaydos, 1998; Leao, Bishop and Evans, 2004; Liu and Feng, 2012), therefore lacking generalisability to wider areas and more variable progression scenarios.

1.4.2 Previous Continental and Global Urban Modelling Approaches

While most urban modelling, not just CA, have been local or regional in focus, there have been some notable efforts of modelling urban features at continental and global extents. They can be broadly separated into two groups: spatially explicit models, which produce spatial maps of the urban feature extents, and non-spatially explicit models, which do not produce spatial maps of the urban features. The non-spatially explicit models instead provide estimates of total urban feature area at some given, often large, spatial scale (Angel *et al.*, 2011; Seto *et al.*, 2011). Here, I will give an overview of the most notable spatially explicit models, highlight why they do not meet the current needs for producing annual estimates of BS for applications involving populations at high spatial resolutions.

1.4.2.1 Spatially Explicit Models

Almost all the spatially explicit models can be Generalised as having two primary components: “demand quantification” and “spatial allocation”. The demand component determines how many target units, within a given larger source unit, should transition from non-urban to urban. The allocation component then determines which of the smaller units should undergo transition in order to produce the final map of urban feature extents.

Tayyebi et al. (2013) created a hierarchical CA-based urban growth model for the coterminous United States which predicted growth from 2001 to 2006 and is summarised in Figure 4.

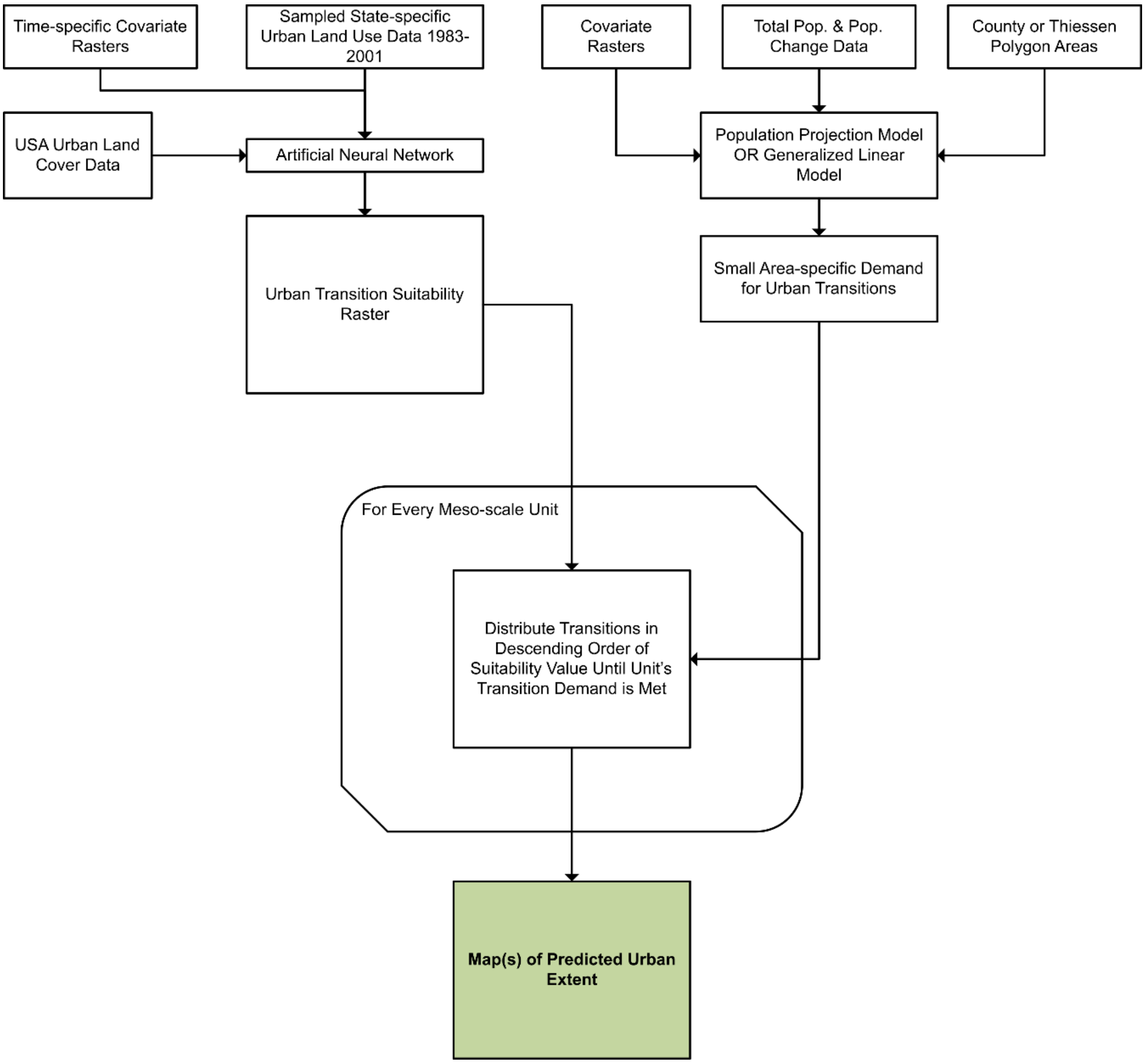

Time-specific Covariate Rasters
Sampled State-specific Urban Land Use Data 1983-2001
Covariate Rasters
Total Pop. & Pop. Change Data
County or Thiessen Polygon Areas
USA Urban Land Cover Data
Artificial Neural Network
Population Projection Model OR Generalized Linear Model
Urban Transition Suitability Raster
Small Area-specific Demand for Urban Transitions
For Every Meso-scale Unit
Distribute Transitions in Descending Order of Suitability Value Until Unit’s Transition Demand is Met
Map(s) of Predicted Urban Extent

Figure 4 Generalised urban thematic land cover model per Tayyebi et al. (2013)

Tayyebi et al. (2013) concluded that urban growth projections driven by local population change or initial local urban quantities and patterns were more suitable and accurate than the application of constant growth rates across the study area. However, they found that the strength of population change in relation to initial urban quantities in predicting urban growth became less clear at small scales (Tayyebi *et al.*, 2013). The approach by Tayyebi et al. (2013) holds a lot of promise, however the reliance on land-use data for training precludes its use across the globe as land use data is often not available or non-existent. Further, the use of an Artificial Neural Network (ANN) requires *a priori* determination of the ANN structure, or structures, to search through before determining the final model. Strengths include the hierarchical framework for distributing urban growth and the allowance for local variation in determining suitability of transition and the amount of transitions across the period.

Linard, Tatem, and Gilbert (2013) created a built land cover growth model for Africa based upon a sample of 40 cities. They trained a Boosted Regression Tree (BRT) on changes in built extent of those cities from 1990 to 2000, as defined by the Atlas of Urban Expansion (Linard, Tatem and Gilbert, 2013). The 40 individual BRTs were combined to then predict for all of Africa (Linard, Tatem and Gilbert, 2013). The modelling process is Generalised in Figure 5.

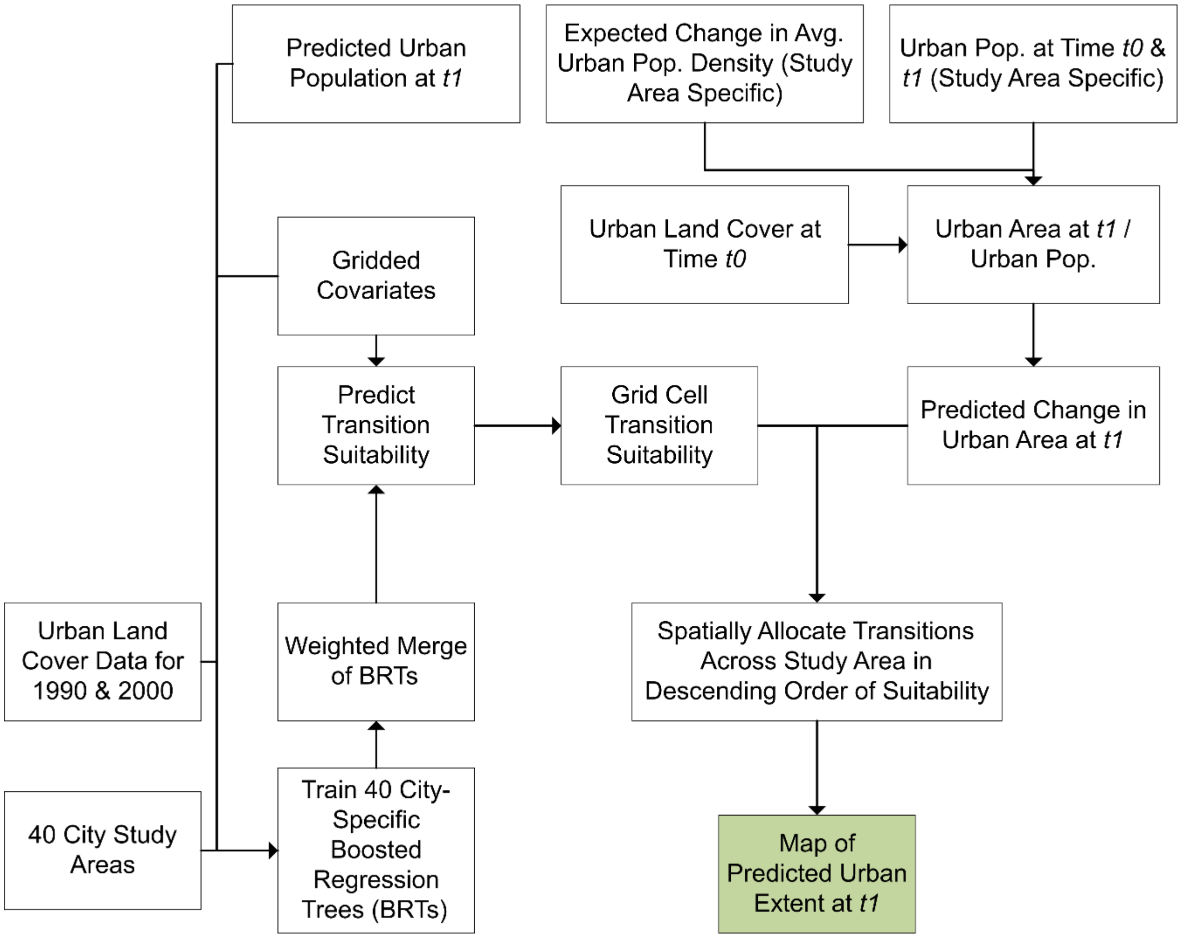
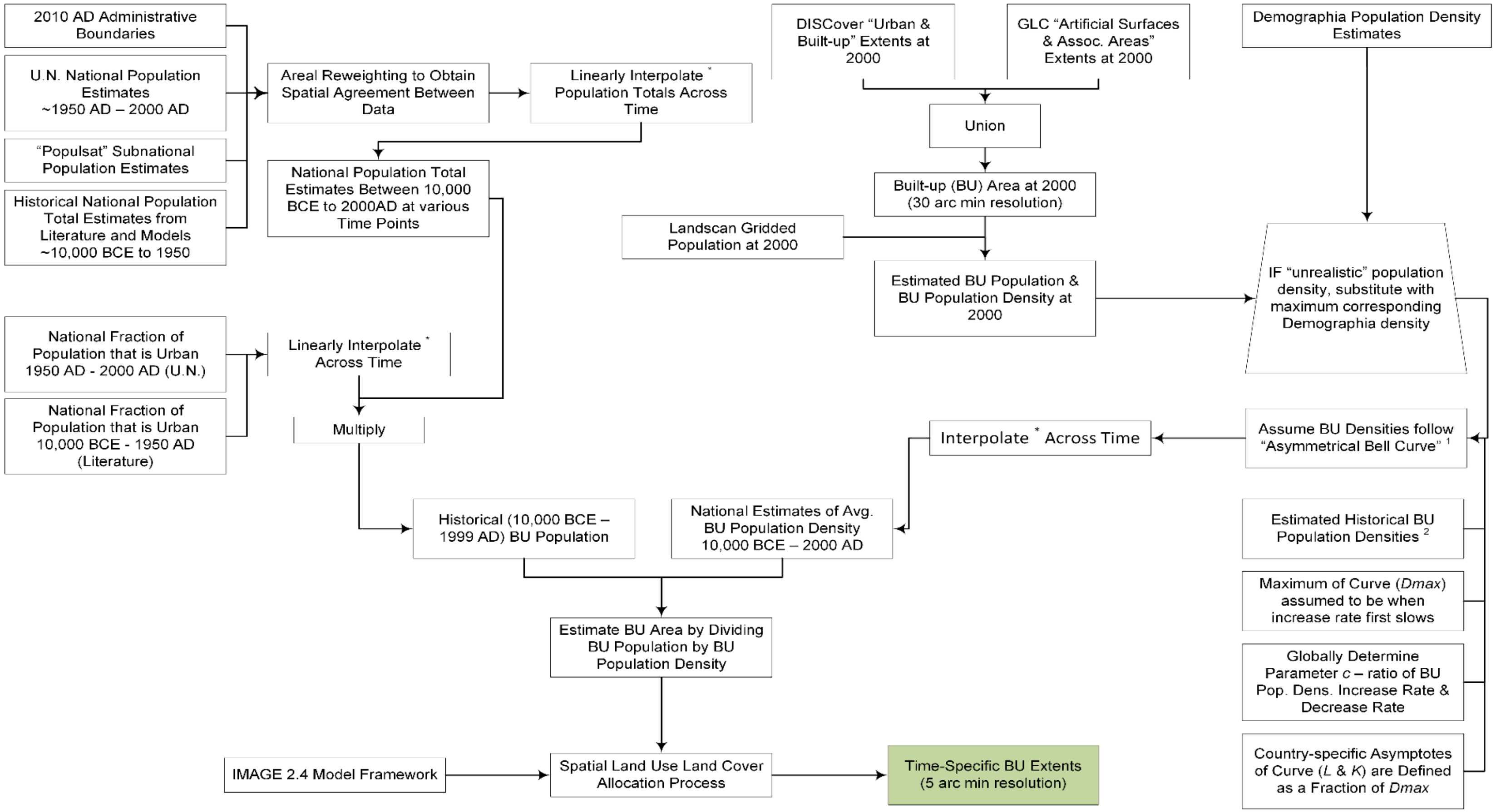

Predicted Urban Population at *t1*
Expected Change in Avg. Urban Pop. Density (Study Area Specific)
Urban Pop. at Time *t0* & *t1* (Study Area Specific)
Urban Land Cover at Time *t0*
Urban Area at *t1* / Urban Pop.
Gridded Covariates
Predict Transition Suitability
Grid Cell Transition Suitability
Predicted Change in Urban Area at *t1*
Urban Land Cover Data for 1990 & 2000
Weighted Merge of BRTs
Spatially Allocate Transitions Across Study Area in Descending Order of Suitability
40 City Study Areas
Train 40 City-Specific Boosted Regression Trees (BRTs)
Map of Predicted Urban Extent at *t1*

Figure 5 Generalised process diagram for the built land cover growth model per Linard, Tatem, and Gilbert (2013).

This model predicted built area growth at approximately 100m x 100m grid cells at ten-year intervals. The transition probability layer allows for local variation in suitability, but limits the demand for “urban” to what equates to a country-, or study area-, wide average (Linard, Tatem and Gilbert, 2013). Additionally, lacking other information at the time, the assumption of a constant linear decrease in population density was made (Linard, Tatem and Gilbert, 2013). Limits aside, Linard, Tatem, and Gilbert (2013) created a continentally applicable urban growth model which relied only on data which could be obtained globally, significantly advancing spatially explicit urban growth modelling and indicating the potential for a global model.

Goldewijk, Beusen, and Janssen (2010) constructed a model whose time coverage was the Holocene, i.e. 10,000 B.C. to 2000 A.D., and was based largely on historical records and anthropological theories to inform model behaviour (Figure 6) (Goldewijk, Beusen and Janssen, 2010). This model predicted urban areas, at approximately 10km x 10km grid cells, based upon historical population density information, fractional land cover percentages, and statistical distributions.

Within this framework (Figure 6), population data varied in both spatial resolution, quality, and source over the study period. Some were historical estimates and some were census or record based, but all were adjusted from their original spatial resolution to contemporary subnational units using simple areal reweighting (Goldewijk, Beusen and Janssen, 2010). Historical urban population densities were derived by fitting country specific normal distributions, the characteristics of which were ascertained by the LandScan-based population densities and the second point where the curve reaches it maximum, i.e. the first time where the rate of urban population decreases for the first time (Goldewijk, Beusen and Janssen, 2010).


2010 AD Administrative Boundaries
U.N. National Population Estimates ~1950 AD – 2000 AD
"Populsat" Subnational Population Estimates
Historical National Population Total Estimates from Literature and Models ~10,000 BCE to 1950
Areal Reweighting to Obtain Spatial Agreement Between Data
Linearly Interpolate * Population Totals Across Time
National Population Total Estimates Between 10,000 BCE to 2000AD at various Time Points
National Fraction of Population that is Urban 1950 AD - 2000 AD (U.N.)
National Fraction of Population that is Urban 10,000 BCE - 1950 AD (Literature)
Linearly Interpolate * Across Time
Multiply
Historical (10,000 BCE – 1999 AD) BU Population
National Estimates of Avg. BU Population Density 10,000 BCE – 2000 AD
Estimate BU Area by Dividing BU Population by BU Population Density
IMAGE 2.4 Model Framework
Spatial Land Use Land Cover Allocation Process
Time-Specific BU Extents (5 arc min resolution)
DISCover "Urban & Built-up" Extents at 2000
GLC "Artificial Surfaces & Assoc. Areas" Extents at 2000
Union
Built-up (BU) Area at 2000 (30 arc min resolution)
Landscan Gridded Population at 2000
Estimated BU Population & BU Population Density at 2000
Demographia Population Density Estimates
IF "unrealistic" population density, substitute with maximum corresponding Demographia density
Interpolate * Across Time
Assume BU Densities follow "Asymmetrical Bell Curve" 1
Estimated Historical BU Population Densities 2
Maximum of Curve (Dmax) assumed to be when increase rate first slows
Globally Determine Parameter c – ratio of BU Pop. Dens. Increase Rate & Decrease Rate
Country-specific Asymptotes of Curve (L & K) are Defined as a Fraction of Dmax

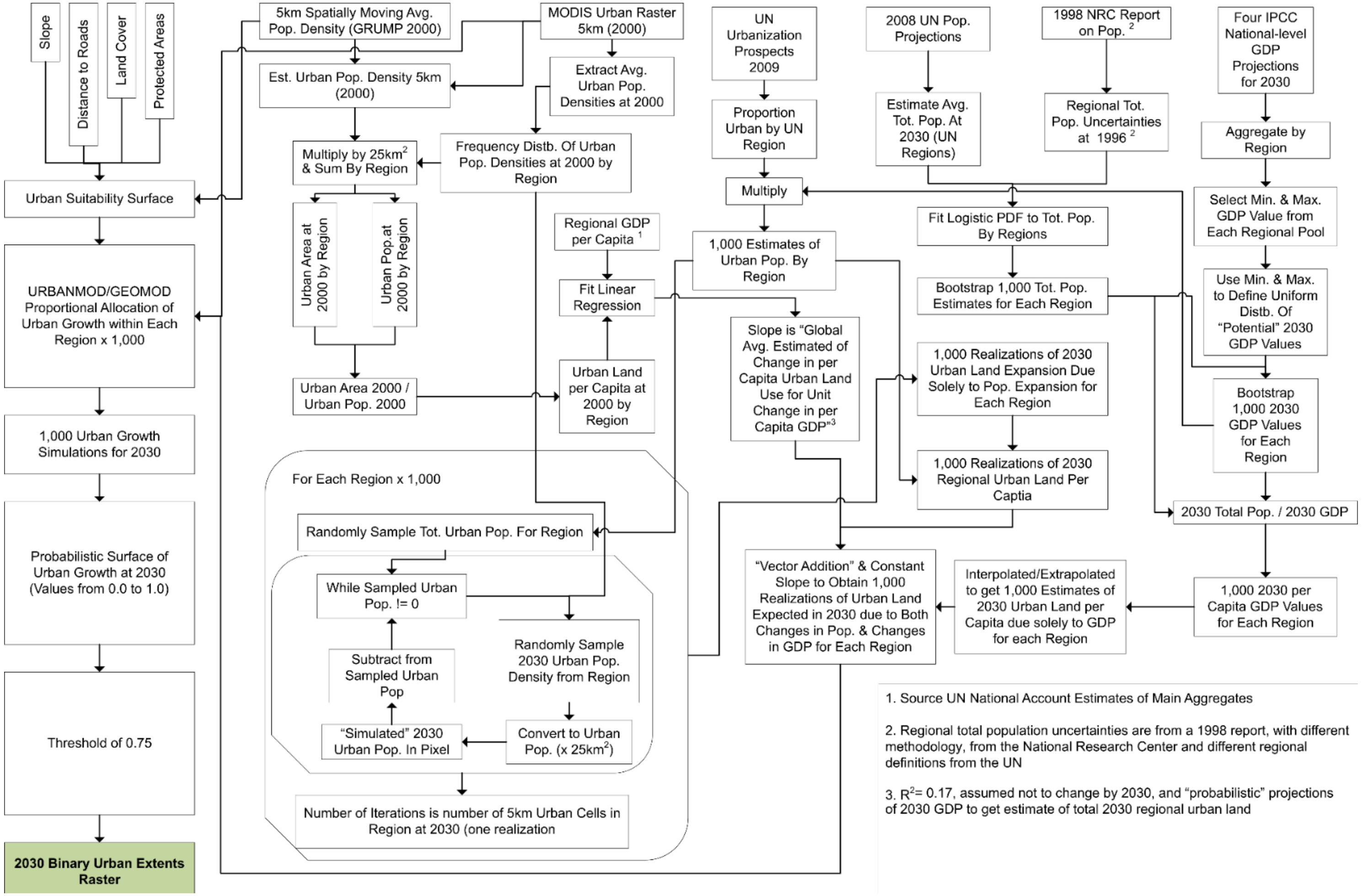
* Paper unclear on method of interpolation. Assumed linear since linear was utilized in other portions of framework
1 Paper unclear on what exactly this means. Could indicate a beta or gamma type curve.
2 Paper unclear as to where these originate. Possible they are derived from the fit curve if authors assume a single maxima that occurs within the modern (1950-2000 AD) data

Figure 6 Generalised process diagram of the settlement model of Goldewijk, Beusen, and Janssen (2010)

Goldewijk, Beusen, and Janssen (2010) note that the uncertainties that accompany historical population data, particularly pre-1700 A.D., are large and the authors even refer to them as "educated guesses." They go further to note that given the subnational units used in the study, this model has artifically low population densities, which are the result of the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984), and is only suitable for country level or regional estimates (Goldewijk, Beusen and Janssen, 2010). While innovative in its sourcing of historical information of population and urbanisation and ambitious in its scope, the applicability of the model outputs across a wide range of spatial scales is a large limiting factor for population mapping and projection below subnational units.

Seto, Guneralp, and Hutyra (2012) constructed a spatially explicit global built land cover growth model, in the supplementary material of their paper, which predicted the built area extents for 2030 at approximately 5km x 5km grid cells based upon a two-step process incorporating population projection and land-use/land-cover change model. The process is Generalised in Figure 7. Seto, Guneralp, and Hutyra (2012) estimated the amount of new built area in 2030 at the U.N. regional level by using distributions of simulated future populations and GDP to obtain 1000 potential realisations of built area land expansion. These predicted growth scenarios allocate the predicted growth within region, at the grid cell level, using a model (URBANMODE/GEOMOD) to distribute the transitions in proportion to the observed distribution of suitability values. That is, if 30 percent of suitability values are 0.4, assumming values range between 0.0 and 1.0, then 30 percent of urban area transitions will be allocated by the model to cells with suitability values of 0.4.

While this effort could be considered the first spatially explicit urban growth model to create future projections on a global extent, unfortunately there are many limits and uncertainty, in conjunction with large spatial, temporal, and methods mismatch. For instance, regional population projections from one dataset, which uses one set of methods and base data, were combined with uncertainty estimates of another dataset, using another set of methods and base data, from 10 years prior whose regional definitions do not match the studies (K C Seto, Guneralp and Hutyra, 2012). Additionally, in estimating the urban expansion due solely to population, the Seto, Guneralp, and Hutyra made the strong assumption that the spatial distribution of population density does not change substantially through time between 2000 and 2030. Lastly, in incorporating both the change in urban land per capita based upon population growth alone and by changes in GDP per capita alone, the model is committing an ecological fallacy by applying the average rate of change of urban land per capita as the GDP per capita changes, i.e. the slope of the regression composed of all regions. Not withstanding the poor fit of the model ($R^2 = 0.17$), the fact that both urban land per capita and GDP per capita are treated as independent measures, which were added as independent vectors, is erroneous; both of these measures have population as their denominator and therefore are mathematically dependent upon population.


Slope
Distance to Roads
Land Cover
Protected Areas
Urban Suitability Surface
URBANMOD/GEOMOD Proportional Allocation of Urban Growth within Each Region x 1,000
1,000 Urban Growth Simulations for 2030
Probabilistic Surface of Urban Growth at 2030 (Values from 0.0 to 1.0)
Threshold of 0.75
2030 Binary Urban Extents Raster
5km Spatially Moving Avg. Pop. Density (GRUMP 2000)
MODIS Urban Raster 5km (2000)
Est. Urban Pop. Density 5km (2000)
Extract Avg. Urban Pop. Densities at 2000
Multiply by 25km² & Sum By Region
Frequency Distb. Of Urban Pop. Densities at 2000 by Region
Urban Area at 2000 by Region
Urban Pop.at 2000 by Region
Urban Area 2000 / Urban Pop. 2000
Regional GDP per Capita 1
Fit Linear Regression
Urban Land per Capita at 2000 by Region
UN Urbanization Prospects 2009
Proportion Urban by UN Region
Multiply
1,000 Estimates of Urban Pop. By Region
Slope is "Global Avg. Estimated of Change in per Capita Urban Land Use for Unit Change in per Capita GDP"3
2008 UN Pop. Projections
Estimate Avg. Tot. Pop. At 2030 (UN Regions)
1998 NRC Report on Pop. 2
Regional Tot. Pop. Uncertainties at 1996 2
Fit Logistic PDF to Tot. Pop. By Regions
Bootstrap 1,000 Tot. Pop. Estimates for Each Region
1,000 Realizations of 2030 Urban Land Expansion Due Solely to Pop. Expansion for Each Region
1,000 Realizations of 2030 Regional Urban Land Per Captia
Four IPCC National-level GDP Projections for 2030
Aggregate by Region
Select Min. & Max. GDP Value from Each Regional Pool
Use Min. & Max. to Define Uniform Distb. Of "Potential" 2030 GDP Values
Bootstrap 1,000 2030 GDP Values for Each Region
2030 Total Pop. / 2030 GDP
1,000 2030 per Capita GDP Values for Each Region
Interpolated/Extrapolated to get 1,000 Estimates of 2030 Urban Land per Capita due solely to GDP for each Region
"Vector Addition" & Constant Slope to Obtain 1,000 Realizations of Urban Land Expected in 2030 due to Both Changes in Pop. & Changes in GDP for Each Region
For Each Region x 1,000
Randomly Sample Tot. Urban Pop. For Region
While Sampled Urban Pop. != 0
Randomly Sample 2030 Urban Pop. Density from Region
Subtract from Sampled Urban Pop
Convert to Urban Pop. (x 25km²)
"Simulated" 2030 Urban Pop. In Pixel
Number of Iterations is number of 5km Urban Cells in Region at 2030 (one realization

1. Source UN National Account Estimates of Main Aggregates

2. Regional total population uncertainties are from a 1998 report, with different methodology, from the National Research Center and different regional definitions from the UN

3. R^2= 0.17, assumed not to change by 2030, and "probabilistic" projections of 2030 GDP to get estimate of total 2030 regional urban land

Figure 7 Generalised process diagram of the built land cover growth model per Seto, Guneralp, and Hutyra (2012)

As I have summarised, the previously constructed continental and global urban growth models were not made with modelling populations at high spatial and or temporal resolution in mind. Large amounts of input “expert opinion”, manually adjusted parameters, or lack of spatial and temporal fineness in either the model variation or the output predicitons leave much opporunity for further innovation. Particularly with regards for leveraging more recent BS data and with the end goal of better population modelling in mind.

1.5 Theme IV: Built Settlement in the Context of Population Modelling

1.5.1 Why Annual Population Modelling is Important

Areal population counts derived from decadal censuses remain a “gold standard” in demography as they are or are the closest data representation of a complete enumeration of populations. However, the quality and coverage of censuses can vary from country to country, can vary within a given country, and can miss or have biases towards informal settlements, remote areas, and mobile populations (Korale, 2002; Sabry, 2010; Tatem and Linard, 2011; Carr-Hill, 2013; Ebenstein and Zhao, 2015; Lucci, Bhatkal and Khan, 2018). More importantly, these decadal census-derived population counts are, by their definition, only measured every 10 years in the best of scenarios, with some countries not having a formal census since the 1960s (Tatem and Linard, 2011; Wardrop *et al.*, 2018). Further, even with the most recent census data, they are often not available at fine spatial scales due to privacy concerns. Here, survey data (Pezzulo *et al.*, 2017), interpolated population counts (Doxsey-Whitfield *et al.*, 2015), and even more novel data, such as mobile phone Call Data Records (Steele *et al.*, 2017; Weber *et al.*, 2018), can partially address the inter-decadal gap. However, their data are not always available, come with their own potential biases, and or do not have entire coverage for a given country or population.

For complete coverage, applications in public health, sustainability, economics, and others still utilise census-derived population counts as the denominator, e.g. mortality rates, CO2 production per capita, gross domestic product per capita, water and food availability (Hay *et al.*, 2004; Tatem *et al.*, 2007; Hanjra and Qureshi, 2010; McDonald *et al.*, 2011; Tatem, 2014; Gibson and Li, 2017). Yet, these areal census-based population data give no information as to the underlying spatial distribution of the actual population, leaving only an incorrect assumption of homogenous population density within a given unit (Tatem *et al.*, 2007).

For instance, in an application looking at rates of malaria prevalence, often given in terms of a rate of "*x* cases per 10,000 people", using the census-based population count at the county level may give an average rate of 3 per 10,000. However, this may hide the fact that there are locations within the county have much higher, and much lower, rates, i.e. the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984). The corresponding spatial distribution of prevalence rates at sub-county levels would also be variable as some function of population. This has important implications as to what locations would or would not be targeted for intervention, e.g. villages in the county with high malaria rates, and spatial distribution of resources for intervention and prevention based upon the sub-county population sizes.

Many have turned to top-down dasymetric disaggregative modelling to, at least partially, address limits of areal population data. The use of dasymetric mapping as applied to human populations was first popularised in 1936 (Wright, 1936). The volume preserving characteristic of dasymetric mapping and its ability to incorporate finer scale supporting information made this an attractive method for disaggregating coarse census-derived population counts to finer spatial scales as the disaggregated counts will always add up to the "gold standard" of the census data. With the advent of GIS, satellite-based remote sensing imagery, the digitisation of more datasets, and the increasing affordability of computational resources, dasymetric mapping of populations has increased in frequency (Bracken and Martin, 1989; Martin, 1989; Flowerdew, Green and Evangelos, 1991; Langford, Maguire and Unwin, 1991; Martin and Bracken, 1991; Deichmann, Balk and Yetman, 2001; Eicher and Brewer, 2001; Mennis, 2003; Balk *et al.*, 2004; Mennis and Hultgren, 2006; Bhaduri, Bright and Coleman, 2007; Cheriyadat *et al.*, 2007; Linard *et al.*, 2010; Gaughan *et al.*, 2013, 2014, 2016; Linard, Gilbert and Tatem, 2013; Sorichetta *et al.*, 2015; Stevens *et al.*, 2015).

These methods transform the irregular area-based census-derived population counts into regularly gridded datasets of population counts by using a weighting layer to redistribute the areal counts to a finer spatial scale. The primary benefit of this procedure being that these datasets give a more realistic and likely representation of the true underlying population distribution within each given unit (Mennis, 2003; Mennis and Hultgren, 2006). Additionally, the gridded data can be subsequently aggregated to different, i.e. non-census, areal units or integrated with other data in different formats for further analysis and application. These gridded population datasets have been used in numerous end applications and research in fields such as sustainability (Gaughan *et al.*, 2019), public health (Linard *et al.* 2010, 2012), and sociology (Pezzulo *et al.*, 2017), to name a few.

When producing gridded population datasets for years that do not correspond to a census or official estimate year, most interpolate or estimate the areal population counts. This is done either using subnational growth rates (Doxsey-Whitfield *et al.*, 2015; Freire *et al.*, 2015), or projecting the pixel-based values using national level urban and rural specific growth rates at subnational levels (Linard *et al.*, 2010; Gaughan *et al.*, 2013; Sorichetta *et al.*, 2015). However, not all of them modify the weighting layer used to disaggregate the interpolated or estimated areal population counts (Gaughan *et al.*, 2013; Sorichetta *et al.*, 2015; Stevens *et al.*, 2015). That is, they grow the areal population, but the built-environment, one of the key markers of human presence on the Earth’s surface (Meyer and Turner, 1992) is not grown with it. Not only does this ignore the relationships between population, the built-environment, and economic activity and maintain static spatial distributions of populations but, if projecting far into the future, could lead to unrealistic population densities.

Estimates of gridded population at annual resolution are needed due to rapidly growing populations in small and medium settlements along with their growing built-environments (Davis, 1965; Montgomery *et al.*, 2003; Cohen, 2006; Angel *et al.*, 2011; Hoalst-Pullen and Patterson, 2011; Acuto, Parnell and Seto, 2018; Zhu *et al.*, 2019). Having settlement data with high spatial resolution, high temporal frequency, and broad temporal coverage is necessary to better understand the spatial distribution of human activities, societal processes, and human environment interactions (Ehrlich, Balk and Sliuzas, 2020). Extremely large amounts of imagery are required to produce a global settlement layer and there is typically a large lag (multiple years) between image acquisition and processing (Esch *et al.*, 2018a; Zhu *et al.*, 2019) and this is further magnified when trying to produce BS features from imagery that are consistent and high in spatial resolution (<= 100m). These rapid changes in both population and settlements require corresponding data and estimates to best inform policy and planning as well as provide better data for understanding the correlates and drivers behind human activities and societal processes (Solecki, Seto and Marcotullio, 2013; United Nations, 2016; Scott and Rajabifard, 2017; Acuto, Parnell and Seto, 2018; Zhu *et al.*, 2019; Ehrlich, Balk and Sliuzas, 2020).

1.6 Aims and Purpose

Given that urban related data is the most important predictor of population density (Nieves *et al.*, 2017), utilising the advances in BS data would appear to present an opportunity towards improving population modelling, which it has already for cross-sectional studies (Freire *et al.*, 2015; Reed *et al.*, 2018; Leyk *et al.*, 2019; Stevens *et al.*, 2020). However, there are gaps in the temporal and spatial continuity of state-of-the-art BS data. Modelling urban related covariates presents an intuitive solution to these BS data issues as well as furthering global population modelling efforts at annual scale while maintaining high spatial resolution. However, existing globally applicable urban models have notable limits: having strict data requirements, requiring much user input or assumptions, and or not allowing for substantial subnational variation. As I have shown, there still exists a need for a globally applicable and spatially explicit urban growth model that:

(i) Only requires globally consistent and available data
(ii) Allows for annual estimates of urban extent
(iii) Requires few assumptions and parameter setting by the user
(iv) Allows for subnational variation
(v) Can interpolate and project into the future based upon observed trajectories, and
(vi) Is able to utilise at and predict at high spatial resolution (<=100m).

Addressing the above would better represent the variation seen in settlement size, type, and growth patterns. Better representing settlement diversity both across space and time can better represent the equally diverse populations that inhabit them. Further, having a flexible modelling framework and globally consistent input data allows for global applicability and comparability across space and time. These characteristics will enable more accurate analyses of past and near future population and BS spatial distributions. This can then serve as a platform for constructing more accurate mid- to long-term population projections and the corresponding settlement footprint on the Earth.

Here, I have developed two modelling frameworks, the Built-Settlement Growth Model – interpolative (BSGMi) and the Built-Settlement Growth Model – extrapolative (BSGMe) that meet these criteria. The following structure of this thesis is as follows. Building upon the concepts introduced with the preceding urban growth models, Chapter 2 presents overviews of key methods and algorithms that were utilised within the BSGMi and BSGMe frameworks. Chapter 3 introduces the BSGMi framework and validates the accuracy of its interpolations in four diverse countries. Chapter 4 introduces the BSGMe framework and validates its extrapolations in four diverse countries. Chapter 5 presents a globally applied validation of the use of the BSGMi outputs for the purposes of population modelling, demonstrating how modelled BS can provide greater information to specific end uses. Chapter 6 presents the larger conclusions and future work based upon Chapters 3-5.

Chapter 2 Automatable, Consistent, and Flexible Modelling Across Large Extents

2.1 Introduction and Background

I created the BSGMi and BSGMe frameworks with the intent of them being globally applicable, meaning they had to be able to run on 249 different countries. Each of those countries was composed of a varying number of subnational units, but globally there were over 14 million subnational units (Doxsey-Whitfield *et al.*, 2015). Given my objective of allowing subnational variation to play a large role in the framework, that meant having over 28 million independent sub-models for each subnational unit within the framework (one model for BS population and one for BS population density). This alone is a large volume of data to handle and is further compounded by the fact I utilised several gridded environmental covariates at 100m spatial resolution, with a single global layer having approximately 51,007,200,000 100m x 100m cells. And while this was a high volume of data storage wise, the breadth of the data was not wide (Wickham, 2014) at all with relatively few covariates for a big data application (Chen, Mao and Liu, 2014). Excluding the BS extents, the only data available with consistent global availability was subnational areal population counts, thematic land cover, temporally static road data, and lights at night data. Even with the few time-specific and (assumed) time-invariant covariates utilised here, the data required 10 Terabytes of storage. Further, each global gridded population map occupied approximately 100 Gigabytes per year. This high-volume data necessitated the use of methods that were computationally efficient, scalable, and largely automatable, i.e. required little manual parameter fitting.

Given these conditions, I adopted a top-down disaggregative framework for both the interpolative and extrapolative Built-Settlement Growth Models (BSGMi and BSGMe, respectively) across both space and time. The algorithms involved in the estimation of the magnitude and timing of the non-BS-to-BS transitions varies within each framework, but all have the commonality of using relative changes in subnational area population to predict BS expansion. Further, the algorithms were automatable and flexible in the data input requirements. This allowed for the capture of a wide variety of data distributions that could be expected when applying the standardised framework over numerous countries, their even more numerous individual subnational units, and to account for the different data scenarios in an interpolative scenario versus an extrapolative one.

Generally, both frameworks can be described as having a “demand quantification” component and a “spatial allocation” component. For the interpolative model (BSGMi), the demand quantification component is informed by logistic growth curves and natural cubic splines for interpolating time-specific built-settlement populations and built-settlement population densities, respectively. The spatial disaggregation component is a dasymetric disaggregation of the predicted built-settlement growth as informed by a random forest (RF) probability of non-BS-to-BS transition layer. In the extrapolative model (BSGMe), the future built-settlement populations and built-settlement population densities are predicted by either an Auto-Regressive Integrated Moving Average (ARIMA) model, and Error Trend Seasonality (ETS) model, or a log-transformed Generalised Linear Model (GLM); whichever of the three exhibited the least error within a rolling origin validation. Rolling origin validations are explained in more detail in Chapter 4. The spatial disaggregation of these demanded transitions is carried out identically to the interpolative model.

Given that relative changes in population are, in part, being used to estimate changes in BS and that my intent is to predict BS to better facilitate population mapping (Chapter 1), concerns of endogeneity are logical. However, as I will describe in Chapter 3 and Chapter 4, the population and BS data are being used at different spatial scales, essentially forming a two-stage hierarchical modelling framework which mitigates this issue. Population is used here at the subnational level to estimate the number and timing of non-BS-to-BS transitions. Population is not used at all in determining where, within each subnational unit, the pixel level non-BS-to-BS transitions occur. Further, there is a large precedence in urban or settlement growth modelling where population is used to determine demand for new at one spatial scale and the spatial allocation of that demand is met at a finer spatial scale without the involvement of population (Sante *et al.*, 2010; Schaldach *et al.*, 2011; K. C. Seto, Guneralp and Hutyra, 2012; Linard, Tatem and Gilbert, 2013; Tayyebi *et al.*, 2013; van Asselen and Verburg, 2013; McKee *et al.*, 2015; Li and Gong, 2016b; Gao and O’Neill, 2019, 2020)

In this chapter, I describe the key structures utilised within both frameworks, dasymetric disaggregation, and the individual algorithmic components utilised within each modelling framework. How these individual components operate, within the larger BSGMi and BSGMe frameworks, and rationale on their choice are covered in Chapters 3 and 4.

2.2 Dasymetric Disaggregation

Dasymetric mapping (or dasymetric disaggregation) is a special case of areal interpolation (Eicher and Brewer, 2001; Mennis, 2003; Mennis and Hultgren, 2006). Areal interpolation is the process of taking a spatial dataset and transforming it from its original areal boundaries to another set of areal boundaries (Mennis, 2003). The original set of areal boundaries are referred to as the “source” areas and the resulting smaller or finer scale areal boundaries are referred to as the “target” areas (Eicher and Brewer, 2001; Mennis, 2003; Mennis and Hultgren, 2006). Areal interpolation disaggregates some attribute value from the source areas to the target areas in a manner where the sum of the disaggregated target values adds back up to their original source area value (Figure 11, top). This feature has led to the adoption of dasymetric mapping techniques for disaggregating census-based population counts (Martin and Bracken, 1991; Bhaduri, Bright and Coleman, 2007; Gaughan *et al.*, 2013; Nagle *et al.*, 2014; Sorichetta *et al.*, 2015; Stevens *et al.*, 2015; Freire *et al.*, 2016; Leyk *et al.*, 2019; Zoraghein and Leyk, 2019). The simplest form of areal interpolation is areal weighting, in which the attribute values are proportionally redistributed from a source area to each target area based upon the proportion of the source area they cover (Mennis, 2003) (Figure 8, top).

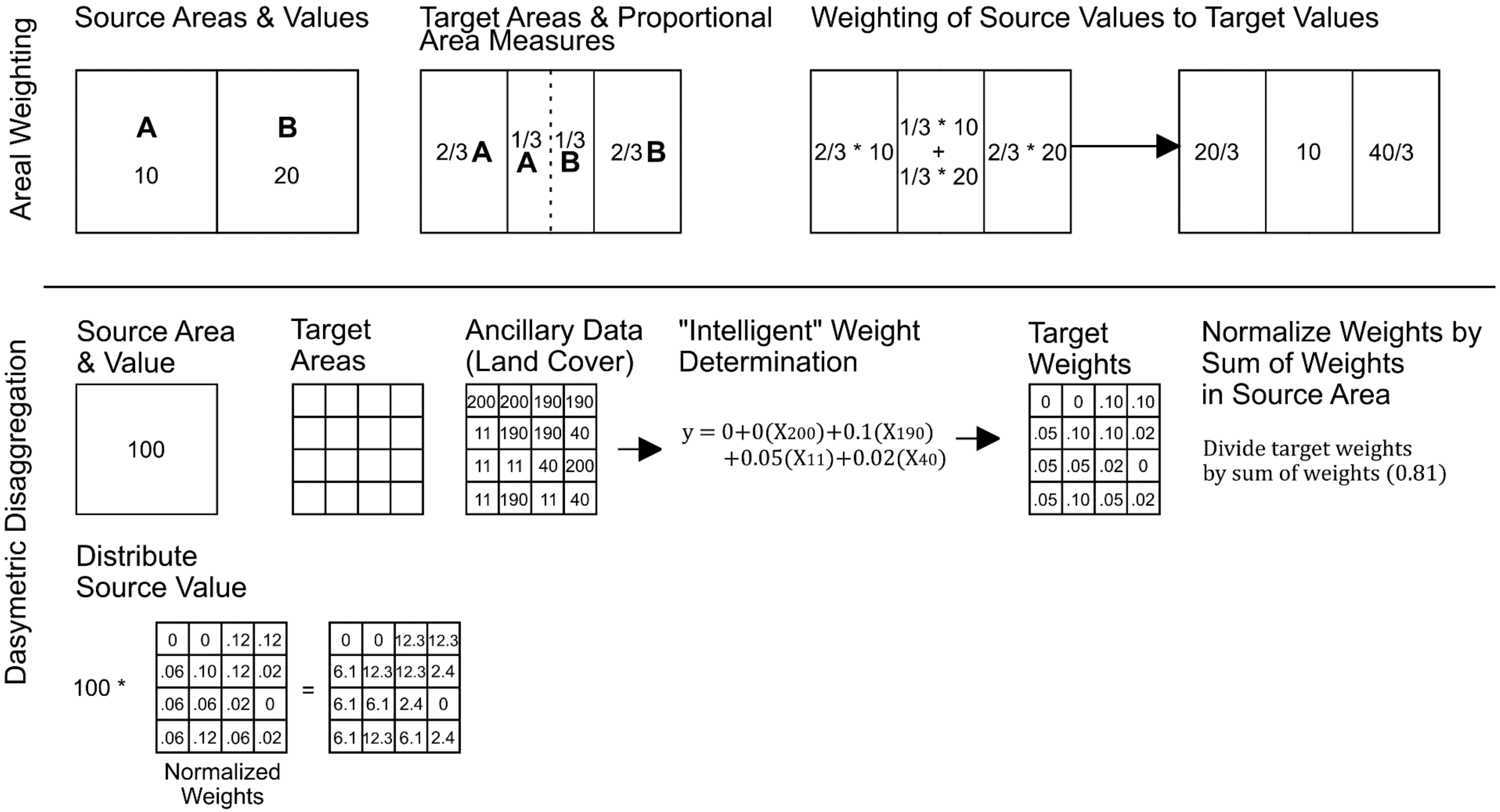


Figure 8 Basic diagram of areal reweighting (top) and dasymetric disaggregation (bottom), with the latter using classified land cover as the ancillary data in determining weights via a multivariate regression.

Mathematically, as adopted from Mennis and Hultgren (2006), this can be stated as:

$$\hat{y}_t = \sum_{s=1}^{n} \frac{y_s A_{s \cap z}}{A_s} \qquad [1]$$

where *s* represents a source zone, *z* represents a target zone, $\hat{y}_t$ is the estimated value of the target area, y_s is the value of the source zone, $A_{s \cap z}$ is the area of intersection between the source and target zone, A_s is the area of the source zone, and *n* is the number of source zones with which *z* overlaps.

Dasymetric disaggregation utilises supporting variable data, known as “ancillary” data sources, at the scale of the target areas to generate weights used to disaggregate the attribute values to the target areas (Eicher and Brewer, 2001; Mennis, 2003). The weights corresponding to the supporting variables are generated by expert knowledge or by statistical relationships between the ancillary data and the attribute being disaggregated (Eicher and Brewer, 2001; Mennis, 2003; Mennis and Hultgren, 2006). In the latter case, this is referred to as “intelligent” dasymetric mapping (Mennis and Hultgren, 2006), with the statistical model determining the weights to be used in the disaggregation (Figure 8, bottom). Mathematically, as adopted from Mennis and Hultgren (2006), this can be described as:

$$\hat{y}_t = y_s \left[\frac{A_t \widehat{D_c}}{\sum_{i \in s} (A_t \widehat{D_c})} \right] \qquad [2]$$

where, given a source area *s* an ancillary area *z* associated with ancillary class *c*, A_t is the area of the target zone, and $\widehat{D_c}$ being the estimated density of the ancillary class *c*. Within intelligent dasymetric mapping, $\widehat{D_c}$ is typically determined by statistical modelling. When the area of all target zones is uniform and identical, as with most gridded data, this equation simplifies to:

$$\hat{y}_t = y_s \left[\frac{\widehat{D_c}}{\sum_{i \in s} (\widehat{D_c})} \right] \qquad [3]$$

While much of the literature has focused on disaggregations across spatial scales, the same concept can work across temporal scales, e.g. disaggregating values from the decadal scale to an annual scale (Zoraghein and Leyk, 2019). The concept of intelligent dasymetric disaggregation is used temporally in the BSGMi framework's "Demand Quantification" component and spatially in the "Spatial Allocation" component. Additionally, it is used in the "Spatial Allocation" component of the BSGMe framework.

2.3 Algorithms Used in Demand Quantification

Here I will give detailed background on the statistical algorithms and curves utilised in the demand quantification components of both the BSGMi and BSGMe frameworks. These algorithms are used in interpolations, generation of dasymetric weights, and extrapolations of values into the near future to estimate demand for non-BS-to-BS transitions at the subnational level.

2.3.1 Built-Settlement Growth Model – interpolative (BSGMi)

An interpolative model is one that predicts within the given range of data observations. The BSGM-interpolative is the modelling framework where, given at two observed time points of data, BS extents are predicted for the years between the given observed time points. This is carried out by interpolating the BS population and the BS population density values of each subnational unit independently.

2.3.1.1 Logistic Growth Curves

Logistic growth curves are widely used and accepted for modelling populations within demography, ecology, and urban modelling (Austin and Brewer, 1971; Wilson, 1976; Ledent, 1982; Cohen, 1995; Smith, 1997). Batty (2009) summarised, "Constrained population growth reflecting both exponential change and capacity which, in turn, reflect densities and congestion are simulated using various kinds of logistic growth." Elaborating on this, Sibly *et al.* (2005) 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." The use of logistic curves in describing this relationship between population density and population growth rates was put forth first by Verhulst (1838). Furthermore, Ledent (1982) showed that urbanisation, the process of population becoming urban, across time can be adequately summarised by "S-shaped curves", specifically the functional logistic form.

If *P(t)* is the total population as a function of time *t*, then the logistic differential equation of total population with respect to time *t* is defined as:

$$\frac{dP(t)}{dt} = rP(t)\left(1 - \frac{P(t)}{K}\right) \qquad [4]$$

where *r* is the intrinsic/per capita growth rate (Sibly, Barker, Denham, Hone and M Pagel, 2005) and *K* is the carrying capacity of the population (Oliver, 1964).The integrated solution of the differential equation becomes:

$$P(t) = \frac{KP_O}{P_O + (K - P_O)e^{-rt}} \qquad [5]$$

where P_o is the total population when *t* = 0. This implies that the integral solution to Equation 5 must be approximately linear when given a fixed value of *K* such that:

$$\ln\left(\frac{P(t)}{K - P(t)}\right) = rt + C \qquad [6]$$

This can be better understood by rearranging Equation 6 to:

$$\frac{\frac{dP(t)}{dt}}{P(t)} = r\left(1 - \frac{P(t)}{K}\right) = r - \frac{rP(t)}{K} \qquad [7]$$

and can be understood to indicate that the proportional rate of population change with respect to time is a linear function, with an average slope of $-\frac{\frac{r(P_n - P_O)}{t}}{K}$ and y-intercept equal to *r*, and decreases as population increases (Oliver, 1964). Equations 5 to 7, can be extended, simply replacing *K*, to include dynamic carrying capacities parameterised on time *K(t)*, allowing for more complex model behaviour (Cohen, 1995; Meyer and Ausubel, 1999a). In applied settings, given a set of *n* known discrete population totals $P_t = \{P_o, \dots, P_n\}$ and discrete carrying capacities $K_t = \{K_o, \dots, K_n\}$ where *t=0,1,…,n*, this means that the value of *r* can be approximated by fitting a linear least squares regression of $\ln\left(\frac{P(t)}{K(t) - P(t)}\right)$ on *t*.

2.3.1.2 Cubic Splines

Splines have been previously utilised for demographic interpolations and forecasts of mortality, fertility, energy demand, and population counts (McNeil, Trussell and Turner, 1977; Ledent, 1982; Booth, 2006; De Jong and Tickle, 2006; Hyndman and Shahid Ullah, 2007; Ugarte *et al.*, 2012; Li *et al.*, 2016). I selected splines as an interpolative method for population density as they maintain agreement with observed values, i.e. the “knots” of a spline (de Boor, 2001). Additionally, splines maintain a smooth rate of change (de Boor, 2001), something to be reasonably expected of population densities barring a catastrophic event such as natural disaster. Further, splines require little additional information other than the data points, allowing for their application to small datasets and requiring no additional inference or assumptions past the degree of the spline’s polynomial (de Boor, 2001).

Given a set of data points (x_i, y_i) of length *n+1*, having a domain from [x_o, x_n], and having a set of points, known as knots, $K = \{x_o, \ldots, x_n\}$ such that $a = x_o < x_1 < \cdots < x_n = b$, the natural cubic spline *S(x)* is a function meeting the following conditions (de Boor, 2001):

1) $S(x) \in C^2[x_o, x_n]$, i.e. *S(x)* has second degree continuity meaning it is twice continuously differentiable across [x_o, x_n]
2) *S(x)* is a third-degree polynomial between each knot [x_{i-1}, x_i], where *i* = 1, …, *n*
3) *S(x)* is an interpolative spline where $S(x_i) = y_i$ for all *i* =0, 1, …, *n*

Conditions 2 and 3 interact where each piecewise section $C_i(x)$ come together to form *S(x)* where:

$$S(x)\begin{cases} C_1(x), x_o \leq x \leq x_1 \\ \qquad \cdots \\ C_i(x), x_{i-1} < x \leq x_i \\ \qquad \cdots \\ C_n(x), x_{n-1} < x \leq x_n \end{cases} \qquad [8]$$

where every $C_i = a_i + b_i x + c_i x^2 + d_i x^3$ with $d_i \neq 0$ for *i* =1, …, *n* (de Boor, 2001). The values of the coefficients are determined by solving a series of derivative equations for each *i* and a “natural” boundary condition is assumed to be known, where the 2nd derivatives of the endpoints are $C_1''(x_o) = C_n''(x_n) = 0$ (de Boor, 2001). This boundary condition implies that as the *x* approaches either endpoint, the curve of the cubic spline approximates a linear function of the form *a + bx*. The resulting spline is an unparameterised curve that produces smooth rates of change and avoids Runge’s Phenomena (Runge, 1901; Epperson, 1987) where, as data is added, the derivatives at each data point increases, resulting in large oscillations of rates of change between data points.

2.3.2 Built-Settlement Growth Model - extrapolative (BSGMe)

An extrapolative model is one that predicts outside of the range of data it was provided with. The BSGM – extrapolative model is given a time series of subnational population count, population density, and BS extent data for observed time periods. The BSGMe then predicts past the last date of observation to provide short-term predictions of annual BS extents.

2.3.2.1 Auto-Regressive Integrated Moving Average (ARIMA) and Error Trend Seasonality (ETS) Models

I utilised ARIMA and ETS models to predict future values of BS population and BS population density at the subnational unit level based upon input time series of values preceding the prediction period. The autoregressive characteristic of ARIMA and ETS models was a primary reason for their use in predicting future populations located in areas of BS and in predicting future BS population density. Limiting the predictive covariates to only be previous values parameterised on time, limits any circular inference that may occur within the derivation of corresponding estimates of BS area.

ARIMA and ETS models are two autoregressive model classes often applied to time series data, sometimes extended to include predictive covariates, but always having dependent model terms based upon preceding values in the observed time series. ETS models are based upon the assumption of non-stationary, i.e. the mean and variance of the underlying process are not constant, and can approximate non-linear processes (Hyndman and Khandakar, 2008). Conversely, ARIMA models assume stationarity and a linear correlation between the values of the time series, but remain a standard statistical benchmark in forecasting on time series (Hyndman and Khandakar, 2008; Fildes and Petropoulos, 2015).

ETS models can be considered a form of exponential smoothing. Exponential smoothing methods have been around since the 1950s and can be described as algorithms that produce point forecasts (Pegels, 1969). These methods were later extended as state-space models, i.e. ETS models. This allowed for the estimation of forecast prediction intervals to accompany the point forecasts, the generation of entire forecast distributions using stochastic processes, and a formal model selection process, giving nine possible model types shown in Table 5. (Ord, Koehler and Snyder, 1997; Hyndman *et al.*, 2002; Hyndman and Khandakar, 2008). ETS models are often Generalised by the form ETS(*E*,*T*,*S*) with *E* representing the error component, *T* representing the trend component, and *S* representing the seasonality component (Hyndman and Khandakar, 2008). Given the phenomenon of my thesis and the short prediction period, I forego utilising ETS models containing seasonal components and only considered non-seasonal ETS models (first column, Table 5).

Table 5 Exponential smoothing methods and corresponding ETS model types. For each shown model there exists two possible variations: one with an additive error component and one with a multiplicative error component. These two variations are indicated by an A or an E prefixed upon the shown abbreviations, e.g. ANN.

	Seasonal Component		
Trend Component	None (N)	Additive (A)	Multiplicative (M)
None (N)	NN	NA	NM
Additive (A)	AN	AA	AM
Additive Damped (A_d)	A_dN	A_dA	A_dM

Each ETS model has an observation equation, describing the relationship between the observations and states, and transition equation(s), which describe the states such as the level, trend, and season and their evolutions across time (Table 6).

Table 6 Recursive and point forecast equations for non-seasonal ETS models. $\hat{y}_{(t+h|t)}$ is the estimated forecast of the time series y at time *t* + *h* given the time series y_t, l_t is the series level at time *t*, b_t is the slope of the series at time *t*, $\alpha, \beta,$ and ϕ are smoothing parameters, and $\phi_h = \phi + \phi^2 + \cdots + \phi^h$.

	Seasonal Component
Trend Component	None
None	$\hat{y}_{(t+h|t)} = l_t$ $l_t = \alpha y_t + (1-\alpha)l_{t-1}$
Additive	$\hat{y}_{(t+h|t)} = l_t + hb_t$ $l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$ $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$
Additive Damped	$\hat{y}_{(t+h|t)} = l_t + \phi_h b_t$ $l_t = \alpha y_t + (1-\alpha)(l_{t-1} + \phi b_{t-1})$ $b_t = \beta(l_t - l_{t-1}) + (1-\beta)\phi b_{t-1}$

For model selection amongst the various ETS models, Hyndman et al. (Hyndman *et al.*, 2002) created an algorithm with the following steps:

1) Apply all appropriate models
2) Optimise parameters using maximum likelihood estimation
3) Select the best model from step one based upon the corrected Akaike Information Criteria (AICc)
4) Calculate point forecasts up to *h* steps past the last observation using the best method
5) Calculate corresponding prediction intervals using either an analytical solution or approximate using bootstrap simulation utilising the underlying state space model (Hyndman and Khandakar, 2008).

Following the above selection procedure, Hyndman et al. (Hyndman *et al.*, 2002) demonstrated that for short-term forecasts (<= 6 time steps), ETS models out performed many other methods on a variety of data sets.

ARIMA models assume that the future values of their specified outcomes are a linear function of current and past observations and current and normally distributed random errors with a mean of zero. They are essentially Auto-Regressive Moving Average (ARMA) models that integrate differencing of the time series to introduce the required condition of stationarity in the time series prior to fitting an ARMA model; hence the "I" in ARIMA. The procedure for determining what specific ARIMA model, the corresponding parameters, and diagnostics was created by Box and Jenkins (Box and Jenkins, 1976) and advanced as a state-space model with a single source of error and an automated fitting procedure by Hyndman et al. (Hyndman *et al.*, 2002). ARIMAs can be Generalised by the form ARIMA(p,d,q)(P,D,Q)$_m$where p is the order (i.e. the number of time-lags) of the autoregressive model, d is the number of differences (i.e. the number of times past values have been subtracted to achieve stationarity), q is the window size or order of the moving average of the model (Hyndman *et al.*, 2002; Hyndman and Khandakar, 2008). P, D, Q, and m parameters are only used for models accounting for seasonality, which, similar to the ETS models, I exclude.

Given a time series of observed data $y_t = (y_1, \dots, y_t) \in \mathbb{R}$, it is sometimes necessary to difference the original series to obtain stationarity, such that the d^{th} difference of y_t can be given as:

$$If\ d = 0: \quad y_t = y_t \tag{9}$$

$$If\ d = 1: \quad y'_t = y_t - y_{t-1} = \mathrm{y_t} - \mathrm{By_t} = (1 - \mathrm{B})\mathrm{y_t} \tag{10}$$

$$If\ d = 2: \quad y''_t = (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) =$$

$$(\mathrm{y_t} - \mathrm{By_t}) - (By_t - B^2 y_t) = (1 - B)^2 y_t \tag{11}$$

where *B* is known as the back-shift operator and in general the *d^th^* order difference can be written as:

$$(1 - B)^d y_t \tag{12}$$

Once a series is stationary, a non-seasonal ARMA(*p, q*) can then be generalised and formally written as (Box and Jenkins, 1976):

$$y_t = c + \phi_t B y_t + \cdots + \phi_p B^p y_t + \theta_t B \varepsilon_t + \cdots + \theta_q B^q \varepsilon_t + \varepsilon_t \tag{13}$$

where c is a constant indicating model drift if greater than zero, ϕ represents optimised weights corresponding to preceding observed values of y_t, and θ represents optimised weights corresponding to preceding error values of ε, which is a gaussian error process with zero mean (Box and Jenkins, 1976; Hyndman and Khandakar, 2008). Hyndman and Khandakar (2008) describe an automated fitting procedure for ARIMAs that relies on unit root tests, iterative step-wise parameter fitting, and selecting the resultant model with the lowest AIC value.

2.3.2.2 Generalised Linear Models (GLMs)

Linear regressions are often used when additional information regarding the shape and distribution of the response in relation to the predictive covariates are unknown. Given that this was the case in the BSGMe, where I had no way of knowing the exact relationship or distribution of BS population counts or BS population density to time, I included a GLM as a potential predictive model. There I regressed log transformed BS population and log transformed BS population density, separately, on a single predictive covariate: time.

Regression-type models associated with Normal, binomial, Poisson, gamma, and other common statistical distributions are composed of systematic and random error components, with all having a linear basis for the systematic component (Nelder and Wedderburn, 1972). Generalised Linear Models (GLMs) provide a single consistent framework for linking the systematic elements of all these regression models with their respective random components through an integrated fitting procedure based upon maximum likelihood, as opposed to the typical least squares method (Nelder and Wedderburn, 1972). GLMs are characterised by:

(i) The random component of the model is from either a Normal distribution or from the exponential family of distributions

(ii) The systematic component is comprised of a set of covariates $x_1,\ldots,$ x_m that produce a linear predictor η defined as $\eta = \sum \beta_i x_i$ where β_i is the fit coefficient for the given covariate x_i

(iii) A link function $g(\cdot)$ that relates the linear predictor to the value of μ in the y datum (McCullagh and Nelder, 1989)

A special, but familiar, case is when the link function $\eta = \mu$, i.e. the "identity" function. This is the simple linear model with errors following a Normal distribution (Nelder and Wedderburn, 1972).

The maximum likelihood fitting procedure is often equivalent to an iterative weighted least-squares procedure with a weight function of:

$$w = \frac{(\frac{d\mu}{dY})^2}{\frac{d\mu}{d\theta}} \qquad [14]$$

with μ being the mean of z and the dependent variable being modified as:

$$y = Y + \frac{z-\mu}{\frac{d\mu}{dY}} \qquad [15]$$

where z are the given observations (Nelder and Wedderburn, 1972). In general, a starting point for the iteration is obtained by approximating $\mu = z$, to then calculate Y, in order to calculate w, set $y = Y$, and obtain the first approximation of the β_i values by regression (Nelder and Wedderburn, 1972). The parameters are then iteratively fit by maximising the likelihood of the given model and measuring the model's deviance from the observed data.

2.4 Algorithms Used in Spatial Allocation

Here I will provide details on the statistical algorithms used in the "Spatial Allocation" components of the BSGMi and BSGMe; namely the random forest (RF), which is used to generate the non-BS-to-BS transition probabilities at the pixel level. Further details on how these probabilities are incorporated and utilised in the BSGMi and BSGMe frameworks are covered in Chapters 3 and 4.

Given that the RF is used to generate the weights used in the dasymetric disaggregation, I will briefly describe and justify the covariates selected for the generation of the weights before describing how a RF is constructed and operates. Details on the data source for these covariates are provided in Chapters 3 and 4. For the RF, I utilised covariates representing accessibility to sizable settlements, elevation and slope, the proportion of areas that was settled within a given radius, distance to various thematic land cover classes, distance to the nearest inland water body, and the distance to the nearest coast.

Travel time or accessibility to settlements above a certain size as a measure of economic "connectedness" has its roots in the "market town" or central business district of Verhulst (1838), Burgess (1925), and Hoyt (1939) and positioning settlements within the more modern polycentric city (Gottman, 1957; Kloosterman and Musterd, 2001, 2001; Parr, 2004; Green, 2007). Spatial proximity to a city or settlement is known to correspond with greater access to infrastructure, services, economic markets, and information (Davis, 1965; Preston and van de Walle, 1978; van Poppel and van der Heijden, 1997; Anas, Arnott and Small, 1998; Dyson, 2011). Additionally, cities have an intertwined relationship with their hinterlands and representing this accessibility as a continuous variable helps move beyond the binary urban-rural dichotomy (Douglass, 1989; Kelly, 1998; Montgomery *et al.*, 2003; Champion and Hugo, 2017; Farrell, 2017). This is somewhat related to the measures of the proportion of an area within a given radius that is settlement. Proximity to existing, and specifically larger, settlements is known to promote infill type growth (Verburg *et al.*, 2002, 2004; Sante *et al.*, 2010; K. C. Seto, Guneralp and Hutyra, 2012; Linard, Tatem and Gilbert, 2013; Tayyebi *et al.*, 2013; Li and Gong, 2016a). Another proximity to existing settlement measure is the distance to nearest settlement edge covariate. Negative values of this covariate correspond to areas inside of settlement agglomerations and are more likely to have infill type growth where settlement does not already exist. Positive values are found outside the edge of settlement agglomerations and would be more likely to have expansion type growth.

Landcovers were included for several reasons. First, they provide contextual information about the settlement via the environment surrounding it (Koning *et* *al.*, 1999; Verburg *et al.*, 1999, 2002; Schaldach *et al.*, 2011; K. C. Seto, Guneralp and Hutyra, 2012; Linard, Tatem and Gilbert, 2013; Tayyebi *et al.*, 2013; van Asselen and Verburg, 2013). For instance, a given area that is settlement with low distances to the nearest cultivated crop and forested landcovers would have a higher probability of being a more rural agricultural settlement which, on average, has potential implications for migration and settlement growth rate (Ledent, 1982; Dyson, 2011). Secondly, different land covers have different costs to convert them from non-BS-to-BS (Verburg *et al.*, 1999, 2002; Schaldach *et al.*, 2011; van Asselen and Verburg, 2013; van Vliet, Eitelberg and Verburg, 2017). For example, converting flat grassland to a suburban housing development takes much less energy and resources than having to cut down and clear a dense forest prior to erecting housing.

Slope and elevation were included for similar reasons. Building settlement on a steep slope, either through sophisticated engineering or re-grading the terrain is much more expensive, i.e. new settlement is less likely to occur, than building on level terrain. Low elevation is correlated with coastal zones, deltas, and river valleys which are all correlated with human settlement, economic activity, and populations (Montgomery *et al.*, 2003; Small and Nicholls, 2003; Small and Cohen, 2004; Small *et al.*, 2018). The latter correlation is why distance to inland water bodies and coasts were included as covariates.

2.4.1 Random Forests

Non-BS-to-BS transitions are a complex phenomenon, exhibiting differences in spatial distribution and environmental context both within and between countries and across time. Compounding this complexity was the high volume of data I had associated with these transitions. Because I required a method which could handle complex, potentially non-linear interactions and was computationally efficient, I selected Random Forests (RFs) for calculating the non-BS-to-BS transition probabilities.

RFs belong to a class of machine learning methods known as “ensemble” methods (Breiman, 2001a). Ensemble methods can be described as having two primary classes: (i) where models of different types are independently created and have their independent predictions combined through some means, such as majority vote and averaging, and, (ii) models of the same type are independently created and then have their independent predictions combined through some means (Chan and Paelinckx, 2008). If all the individual models have slightly better than random accuracy, i.e. a “weak learner”, the aggregate prediction of the models is, on average, better than any single model (Schapire, 2013).

RFs are a non-parametric modelling method composed of hundreds of Classification And Regression Trees (CARTS) (Figure 9). CARTs recursively subdivided each “node”, or set, of data based upon some splitting criteria, typically a logical statement or inequality that maximises the homogeneity or “information gain” of the resulting smaller sub-nodes (data subsets). RFs are robust to noisy data, can handle small and large data sets, able to capture non-linear phenomena and complex interactions, and handle both categorical and numeric data (Breiman, 2001a; Liaw and Wiener, 2002).

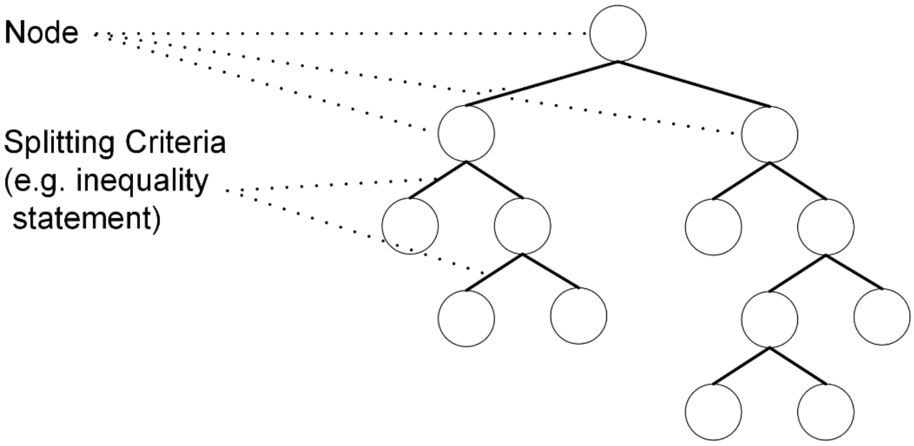


Figure 9 Generalised diagram of a Classification And Regression Tree (CART).

Further, RFs require almost no manual parameter setting, are highly efficient and parallelisable, and are extremely robust to overfitting; more so than other popular methods such as boosted regression trees, and artificial neural networks, and support vector machines (Breiman, 2001a). RFs have also been shown in at least one previous study to outperform SVMs in predicting non-BS-to-BS transition (Kamusoko and Gamba, 2015).

RFs are set apart from other CART and ensemble methods using CARTs by two key characteristics: (i) the use of bagging to select training sets for individual CART construction, and, (ii) the random selection of a covariate (or a linear combination of several random covariates) to use as splitting criteria at each node. Bagging, or “bootstrap aggregating”, is a procedure in which the total data available for training is sampled with replacement to create a training set for constructing a given tree (Figure 10) (Breiman, 1996, 2001a).

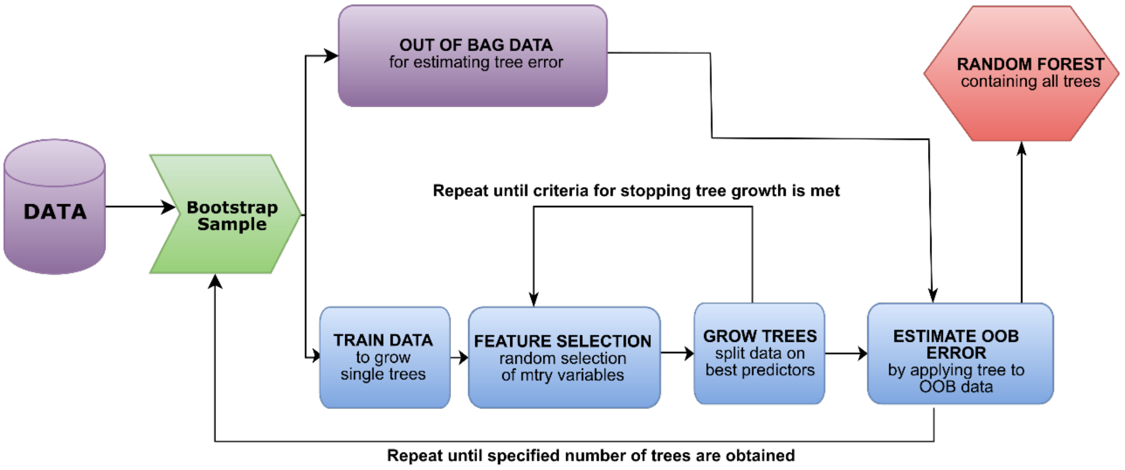

OUT OF BAG DATA
for estimating tree error
RANDOM FOREST
containing all trees
DATA
Bootstrap Sample
Repeat until criteria for stopping tree growth is met
TRAIN DATA
to grow single trees
FEATURE SELECTION
random selection of mtry variables
GROW TREES
split data on best predictors
ESTIMATE OOB ERROR
by applying tree to OOB data
Repeat until specified number of trees are obtained

Figure 10 General diagram of how a random forest is constructed

Typically, a sample equal in size to 2/3 of the total available data is taken, with the unsampled observations being called the “Out-Of-Bag” (OOB) sample (Breiman, 2001a; Liaw and Wiener, 2002). A CART is then created using the bagged sample with a random sample of the total covariate space evaluated for use as splitting criteria at each node (Breiman, 2001a). The splitting criteria, defined by a single covariate value inequality or as an inequality of a linear combination of several covariates, is determined by maximising the node purity of the two sub-nodes resulting from the split (Figure 10) (Breiman, 2001a; Liaw and Wiener, 2002). Within a classification RF, node purity is measured by the Gini impurity; a measure of the probability of a random element being classified incorrectly if randomly classified by a given node’s distribution of classes (Breiman, 2001a). After an individual tree is constructed, the OOB sample is given to the tree to predict upon and the error of the tree on the OOB sample is recorded. Aggregating the OOB error of the individual trees gives the generalised error of the entire RF model and serves as an unbiased internal cross-validation (Breiman, 2001a).

While RFs are a “black-box” method, due to the fact that hundreds of individually interpretable CARTs are not interpretable as a whole, they do provide a measure of covariate importances (Breiman, 2001a). When used for regression, a RF will take the training data and, for a given covariate of interest, randomly permute the covariate data across the rows of data, breaking the association of the covariate with the outcome (Breiman, 2001a). The errors of the model are re-evaluated and the Percent Increase in the Mean Square Error (PER.INC.MSE) of the model, when compared to the intact data, is assessed (Breiman, 2001a). A higher PER.INC.MSE indicates a more important covariate (Breiman, 2001a).

2.5 Balancing Automation with Reliability and Transparency

There is a trade-off between the automatability and scalability of methods against the interpretability of methods and intuitive measures of the methods' consistency and reliability. Within my framework, where I am fitting models independently for each subnational unit resulting in the training tens of millions of independent models , the question naturally arises as to how accurate, reliable, and consistent the modelling framework, and its constituent methods, can be. Here, I will briefly discuss some of the practices and considerations that allow for the modelling framework to maintain usability and interpretability.

2.5.1 Documenting framework output for assessment

One of the key strengths of programmable modelling is its ability to self-document methods through the code and the comments embedded in the code. This lends itself to transparency and replicability and can serve as a first point for understanding how the model framework operated. Further, within both the BSGMi and the BSGMe modelling framework, I programmed the code to save the fit random forest model objects which contain all the information regarding the parameters used in the fitting of the model, the individual tree structures, and the internal cross-validation measures (Liaw and Wiener, 2002). These model objects allow for the replication of model predictions, contain details regarding how covariates contribute to the model structure, and contain the internal error estimates of the model. Ultimately, preserving model objects promotes transparency and replicability.

For each country that is run through the BSGMe framework, information allowing for the reconstruction of the subnational unit level ARIMA, ETS, or GLM model that is selected for prediction is recorded in a data table. This allows users to assess the type of model fit, the model parameters, and the model fit error against the training data during the rolling origin validation procedure. Much of this recorded metadata is not easily human interpretable, solely due to the number of records. However, should a user have a query regarding individual subnational units, the metadata is quite interpretable. A larger strength is that the metadata records are consistently and programmatically created. This means that they can be subsequently used for further analyses and diagnostics. Further work may look at an interactive dashboard for greater manual human access and interpretation.

One might naturally ask if the provided measures of uncertainty are of the final model output, i.e. the spatially explicit settlement extents. While the algorithms utilised within the modelling framework can produce measures of uncertainty, the dasymetric disaggregation procedure, which uses those predictions as relative weights within a given unit, precludes the measurement of the propagation of those uncertainties to the final disaggregated values and spatial extents. This is a noted limit of this top-down approach (Mennis, 2003; Mennis and Hultgren, 2006; Nagle *et al.*, 2014; Stevens *et al.*, 2015). However, the level at which the uncertainty estimates are provided are still valid. For instance, within the BSGMe framework and for some subnational unit and year, say there is an uncertainty estimate given based upon its fit ARIMA model predicting the area to transition from non-BS to BS. The uncertainty of the amount of area to transition at the subnational level is valid.

While the random forest models in the top-down framework is capable of producing uncertainty estimates for its predicted pixel-level population densities, because those predicted population densities are then used as relative weights within each subnational unit, the uncertainty estimates do not propagate (Mennis and Hultgren, 2006; Nagle *et al.*, 2014; Stevens *et al.*, 2015). This means that uncertainty or error estimates must come from *post hoc* comparison to a finer scale (than the subnational unit values were disaggregated from) independent dataset. Unfortunately, these independent and time specific finer scale population or BS data do not often exist (Sinha *et al.*, 2019). However, as a rule of thumb, greater uncertainty can be expected in areas and time periods where the source area/time period is much larger than the final disaggregated scale. Additionally, it would be logical to expect that a disaggregative model whose weight constructing model does not capture a high proportion of the data variance at the source level, e.g. low r2 value, would not be expected to provide good relative weights at the target level (Sinha *et al.*, 2019). However, because the weight constructing model is typically trained at a higher spatial/temporal scale than it is predicting, i.e. it is committing an ecological fallacy, high or low variance explained at the source level is not a guarantee of a good or accurate disaggregation (Sinha *et al.*, 2019).

While this recording of model objects and metadata is important for replicability and transparency in science, they are not the focus of this thesis. I will now introduce and validate the BSGMi and BSGMe frameworks and quantify how they can contribute to population mapping efforts.

2.6 Guidance on Reading of the BSGM Works

Given that this is thesis is in a three-paper format, some of the articles (Chapter 3 and Chapter 4) are already published. Rather than modify them from their peer-reviewed format, they are placed within this larger thesis as is. That being said, the ends of Chapter 3 and Chapter 4 have their own reference sections and accordingly any references in Chapter 3 and Chapter 4 should refer to these reference sections rather than the larger Bibliography at the end of this thesis. Chapter 5 is the exception, still being in the review process, and any references in Chapter 5 refer to the Bibliography.

It is also worth noting to the reader, that Chapter 5 uses the BSGMi alpha version (BSGMiα), which uses exponential growth/decay curves to interpolate the BS population and BS population density. This was an early version of the BSGMi, which is validated in Chapter 3, that was produced at an accelerated rate to meet the project constraints of the WorldPop Global project (see Preface). The benefit of having a working version of the model early, meant that the Global project carried out modelling using the framework across 249 countries between 2000 and 2014. However, there are some key differences between the BSGMiα and the BSGMi presented in Chapter 3. The BSGMi uses a logistic curve for interpolating BS population and natural cubic splines for interpolating BS population density as opposed to the aforementioned exponential growth in the BSGMiα. Additionally, the natural cubic splines of the BSGMi are fit across all observed BS population density points, 2000, 2005, 2010, 2015 in the case of Chapter 3, as opposed to only between two points within the BSGMi α. A brief discussion as to how these differences may manifest in the predicted settlement extents are provided in Appendix B.

Chapter 3 Annually modelling built-settlements between remotely-sensed observations using relative changes in subnational populations and lights at night

Nieves, J. J., Sorichetta, A., Linard, C., Bondarenko, M., Steele, J. E., Stevens, F. R., Gaughan, A. E., Carioli, A., Clarke, D. J., Esch, T., & A. J. Tatem. "Annually modelling built-settlements between remotely-sensed observations using relative changes in subnational populations and lights at night"

The version included in this thesis is the author's version of a work that was accepted for publication in *Computers, Environment and Urban Systems*. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in *Computers, Environment and Urban Systems*, 80, 2020.

https://doi.org/10.1016/j.compenvurbsys.2019.101444

ABSTRACT

Mapping urban features/human built-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 features/human built-settlement datasets have become available, issues still exist in remotely-sensed imagery due to spatial and temporal coverage, adverse atmospheric conditions, and expenses involved in producing such datasets. Remotely-sensed annual time-series of urban/built-settlement extents therefore do not yet exist and cover more than specific local areas or city-based regions. Moreover, while a few high-resolution global datasets of urban/built-settlement extents exist for key years, the observed date often deviates many years from the assigned one. These challenges make it difficult to increase temporal coverage while maintaining high fidelity in the spatial resolution. Here we describe 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 built-settlement datasets in four test countries located in 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 built-settlement in each year. This modelling framework shows strong promise for filling gaps in cross-sectional urban features/built-settlement datasets derived from remotely-sensed imagery, provides a base upon which to create urban future/built-settlement extent projections, and enables further exploration of the relationships between urban/built-settlement area and population dynamics.

Keywords

Built-settlements, urban features, spatial growth, random forest, dasymetric modelling, population

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 expected to occur within Africa and Asia (Angel, Sheppard and Civco, 2005)(Angel, Sheppard, & Civco, 2005; United Nations, 2015b). Further, rapidly changing magnitudes and distributions of both built-settlements and populations have significant implications for sustainability (Cohen, 2006), climate change (McGranahan, Balk, & Anderson, 2007; Stephenson, Newman, & Mayhew, 2010), and public health (Chongsuvivatwong et al., 2011; Dhingra et al., 2016), 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 classified landscape types (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) (United Nations, 2016).

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 (United Nations, 2015a, 2018). While not conducive to applications requiring global consistency in definitions (D Potere & Schneider, 2007), 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” (Merriam-Webster, 2019) is a complex amalgamation of the physical environment, population, economics, movements, and connectivity (Angel, Parent, Civco, Blei, & Potere, 2011; Berechman & Gordon, 1986; Burgess, 1925; Cohen, 2004, 2006; Dyson, 2011; Gottman, 1957; Haas, 2010; Harris & Ullman, 1945; Hoyt, 1939; Ledent, 1982; W. B. Meyer & Turner, 1992; Parr, 2004; Pozzi & Small, 2005; Schneider, Friedl, & Potere, 2010; Seto, Fragkias, Guneralp, & Reilly, 2011; Southworth, 1995; Von Thunen, 1966; Zelinsky, 1971). Figure 1, Part A gives a Generalised diagrammatic view of the factors contributing to the concept of urban.

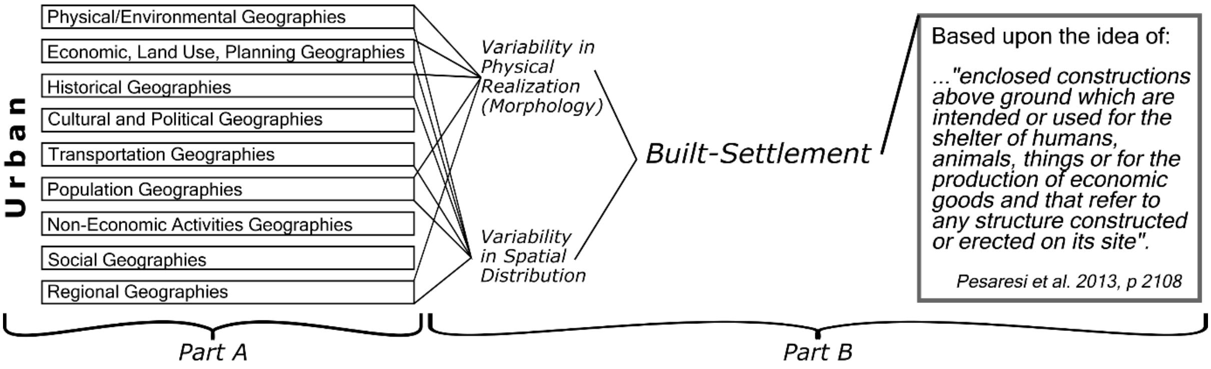

Urban
Physical/Environmental Geographies
Economic, Land Use, Planning Geographies
Historical Geographies
Cultural and Political Geographies
Transportation Geographies
Population Geographies
Non-Economic Activities Geographies
Social Geographies
Regional Geographies
Variability in Physical Realization (Morphology)
Variability in Spatial Distribution
Built-Settlement
Based upon the idea of:
..."enclosed constructions above ground which are intended or used 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".
Pesaresi et al. 2013, p 2108
Part A
Part B

Figure 1. Generalised concept of "urban" (Part A), the conceptual relations and definition of "built-settlement" (Part B) as related to urban, 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 (RS) 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 the types of materials used, differences in urban morphology, and the surrounding environmental context (Schneider et al., 2010; Schneider & Woodcock, 2008; Small, 2009). Initially, remotely sensed urban definitions were optically-based thematic classifications of land cover, typically capturing the "built-environment," including buildings, roads, runways, and, sometimes erroneously, bare soil (Bartholomé & Belward, 2005; David Potere, Schneider, Angel, & Civco, 2009; Schneider, Friedl, McIver, & Woodcock, 2003; Schneider et al., 2010). Other definitions have utilised urban delineated extents and densities based upon Lights-At-Night (LAN) data (Elvidge, Baugh, Kihn, Kroehl, & Davis, 1997; Henderson, Yeh, Gong, Elvidge, & Baugh, 2003; Xue Liu, de Sherbinin, & Zhan, 2019; Shi et al., 2014; Small, Elvidge, Balk, & Montgomery, 2011; Small, Pozzi, & Elvidge, 2005; Wicht & Kuffer, 2019). Later improvements using supporting information about the surrounding environment and vegetation during post-processing helped better discern the true built-environment from the surrounding land covers (Schneider et al., 2010). Other notable advances include the use of high resolution orthographic imagery to detect subtle short-term built-environment change in China (X Huang, Wen, Li, & Qin, 2017) and the use of Landsat imagery to create multi-temporal thematic representations of the built environment across the globe (Xiaoping Liu et al., 2018).

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 (Pesaresi et al., 2016, 2013) or utilising 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 (i.e. built-settlements), while attempting to exclude other anthropogenic land covers (Esch et al., 2013). However, it remains a challenge to produce consistent global urban feature/built-settlement products while maintaining high temporal and spatial fidelity, meaning most of the global multi-temporal urban feature/built-settlement data sets refer to few time points across a larger time period. Further, the cost of producing these data remains relatively high (Esch et al., 2018a) 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 all imagery are suitable for producing high-frequency global urban feature/built-settlement data sets. This is either because of the aforementioned reasons and/or because the processing cost may not be viable or the funds for such endeavours may not be available.

One way to address these issues is to leverage years where RS-based urban feature/built-settlement extractions with high spatial fidelity are available and interpolate for missing time points and areas of interest by modelling between available years. Overall, urban feature/built-settlement growth models have disproportionately focused on high-income countries, which have different dynamics than low- and middle-income countries (Angel et al., 2005; Linard, Tatem, & Gilbert, 2013; Seto et al., 2011; United Nations, 2015b), and most have been limited to city or regionally specific models (Barredo, Demicheli, Lavalle, Kasanko, & McCormick, 2004; Batty & Xie, 1994; Clarke & Gaydos, 1998; Clarke, Hoppen, & Gaydos, 1997; Xin Huang, Hu, Li, & Wang, 2018; Leao, Bishop, & Evans, 2004; Linard et al., 2013; Sante, Garcia, Miranda, & Crecente, 2010; White & Engelen, 1997, 2000). Previous methods of modelling urban feature/built-settlement growth across space and time at the continental and global scales include land cover/land use transition models (Tayyebi et al., 2013; Verburg, Schot, Dijst, & Veldkamp, 2004) and cellular automata models (Batty, 2009; Sante et al., 2010; Verburg et al., 2002), with features or thematic classes extracted from remotely-sensed imagery being the primary source of cross-sectional input for these models (Esch et al., 2013; Patel et al., 2015; Pesaresi et al., 2016, 2013;

Schneider et al., 2010). Readers are referred to Li & Gong (2016) and Sante et al. (2010) for comprehensive reviews of the wide field of cellular automata models as applied to urban feature/built-settlement growth modelling. Of the few models predicting urban feature/built-settlement growth across the globe within a standardised framework, almost none provided explicit spatial prediction finer than country level summaries (Angel et al., 2011; Seto et al., 2011). Global models that did provide explicit spatial predictions, did not allow local sub-national variations to drive the modelled changes or had not been assessed against comparable existing datasets (Angel et al., 2011; Goldewijk, Beusen, & Janssen, 2010; Linard et al., 2013; Seto, Guneralp, & Hutyra, 2012).

Building upon the previous work of these models, in this study, we leveraged the recently available multi-temporal global urban feature/built-settlement datasets, global environmental datasets, subnational census-based population data, and computational methods to develop a flexible globally applicable modelling framework based upon random forest classification trees, population growth curves, and cubic splines. Our specific objectives were to i) determine if random forests can reasonably predict the probability of non-BS to BS transition probabilities, ii) use the predicted surface of non-BS-to-BS transition probabilities as input to an automated framework to annually estimate spatially explicit BS extents using sub-nationally driven geospatial covariates and population counts, iii) validate the model performance and validate the model outputs.

Because the focus of this study is on modelling urban feature/built-settlement extents that better represent where people may be located, we adopted the Global Human Settlement Layer (GHSL) concept of “built-settlement” (BS) (Figure 1), 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.” (Pesaresi et al., 2013, p. 2013). We further Generalised the definition of BS to include other datasets that attempt 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 “urbanisation” is taken within a remote sensing context to be the physical transition from a non-BS area to a BS area.

2. METHODS AND DATA

2.1 Study Areas

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 detail of input census-based population data, as measured by the average spatial resolution (Tobler, Deichmann, Gottsegen, & Maloy, 1997). The countries selected here were Panama, Switzerland, Uganda, and Vietnam, which are located in rather contrasting geographies and environmental/urban biomes (Schneider et al., 2010) and represent quite different cultural and developmental contexts (from low-, middle-, to high-income countries). While it is known that many urban feature datasets have difficulty classifying the built-environment in arid regions, this is more a concern of the selected representation of BS input into the modelling framework rather than an issue for the framework itself; an inaccurate or "noisy" input will always produce poor results in an interpolative model.

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)	Period Transition Prevalence
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, and can be thought of as analogous to pixel resolution with smaller values indicating finer areal data and vice versa (Tobler et al., 1997)

2.2 Built-settlement Data

We chose to use the "Urban areas" thematic class, class 190, from the ESA CCI land cover 300m annual global land cover time-series from 1992 to 2015 dataset (https://www.esa-landcover-cci.org/; hereafter ESA) for our study. It was selected for its annual coverage, allowing for the withholding of years in the model training process for validation of latter modelled outputs. For our period of interest, 2000 to 2015, the ESA time-series includes annual 10 arc sec resolution (~300m at Equator) datasets produced by looking for thematic class changes from a baseline land cover map, obtained using MERIS imagery, using 30 arc second (~1 km at the Equator) SPOT VGT imagery (1999-2013) and PROBA-V imagery (2014-2015) (UCL Geomatics, 2017). 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 (UCL Geomatics, 2017). To reduce false detections, changes must be observed over two years or more (UCL Geomatics, 2017). Furthermore, the GHSL (Pesaresi et al., 2016, 2013) and Global Urban Footprint (GUF) (Esch et al., 2013) datasets are utilised in defining the extents of the ESA "Urban areas" class (UCL Geomatics, 2017), 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 areas" class user and producer accuracies between 86-88 percent and 51-60 percent, respectively (UCL Geomatics, 2017). We also tested and validated a single year, 2010 as predicted from the years 2000 and 2015, from an alpha version of the forthcoming multi-temporal World Settlement Footprint (WSF) dataset, known as WSF Evolution (Esch et al., 2018a), and present the results in the Supplementary Material.

2.3 Population Data

Annual population counts from 2000 to 2015 for subnational areas were provided by the Center for International Earth Science Information Network (CIESIN) in tabular format with unique IDs corresponding to unique subnational unit IDs (Doxsey-Whitfield et al., 2015). Populations and areas of the 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. (2015) for the interpolation and extrapolation of population between 2000 and 2015, inclusive, using years of official counts or estimates. The level of spatial fineness of the subnational units varies from country to country. Typically, all countries are at level 2 or finer, with some countries, such as the USA, being at the block level (level 5) (Doxsey-Whitfield *et al.*, 2015).

2.4 Geospatial Data

We selected a number of covariates based upon previous urban feature/built-environment models (Linard et al., 2013; Verburg, de Koning, Kok, Veldkamp, & Bouma, 1999; Verburg et al., 2002) to give the model information on the immediate environmental/land cover context and connectivity of urban feature/built-settlements. Ultimately, the model is not dependent on any specific geospatial covariates, retaining a level of flexibility for use in a wide variety of applications. For example, a minimal set of globally available predictive covariates to produces inputs for other modelling efforts while avoiding potential issues relating to endogeneity. In the case presented here, annually available covariates, or single time point covariates reasonably assumed to be time invariant, were used either in the direct calculation of transition probabilities or in the remainder of the disaggregative process (Table 2). As detailed in Lloyd et al. (2019), all covariates were pre-processed, appropriately resampled, and matched 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 of the covariates (Table 2) and the ESA 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 Appendices.

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

Covariate	Variable Name(s) in Random Forest	Description	Use[b,d]	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	(ESA CCI, 2017)
DTE Built-settlement	esa_cls190_dst_<*year*>	Distance to the nearest BS edge	Spatial Allocation[d]	2000	10 arc sec	(ESA CCI, 2017)
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[d]	2000	10 arc sec	(ESA CCI, 2017)
Elevation	Topo	Elevation of terrain	Spatial Allocation[d]	2000 – Time Invariant	3 arc seconds	(Lehner, Verdin, & Jarvis, 2008)
Slope	Slope	Slope of terrain	Spatial Allocation[d]	2000 – Time Invariant	3 arc seconds	(Lehner et al., 2008)
DTE Protected Areas Category 1	wdpa_cat1_dst_2015	Distance to the nearest level 1 protected area edge	Spatial Allocation[d]	2015	Vector	(U.N. Enviroment Programme World Conservation Monitoring Centre & IUCN World Commission on Protected Areas, 2015)
Water	---	Areas of water to restrict areas of model prediction	Restrictive Mask		5 arc second	(Lamarche et al., 2017)
Subnational Population	---	Annual population by sub-national units	Demand Quantification	2000 -2020, annually	Vector	(Doxsey-Whitfield et al., 2015)
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 (WorldPop, Department of Geography and Geosciences, Département de Géographie, & Center for International Earth Science Information Network (CIESIN), 2018; Zhang, Pandey, & Seto, 2016) VIIRS(Earth Observation Group NOAA National Geophysical Data Center, 2016; WorldPop et al., 2018)
Travel Time 50k	tt50k	Travel time to the nearest city centre containing at least 50,000 people	Spatial Allocation[d]	2000	30 arc second	(Nelson, 2008)
Urban Accessibility 2015	urbanaccessibility_2015	Travel time to the nearest city edge	Spatial Allocation[d]	2015	30 arc second	(Weiss et al., 2018)
ESA CCI Land Cover (LC) Class [a]	ccilc_dst<*class number*>_<*year*>	Distance to nearest edge of individual land cover classes	Spatial Allocation[d]	2000	10 arc second	(ESA CCI, 2017)

Covariate	Variable Name(s) in Random Forest	Description	Use[b,d]	Time Point(s)	Original Spatial Resolution(s)	Data Source(s)
Distance to OpenStreet Map (OSM) Rivers	osmriv_dst	Distance to nearest OSM river feature	Spatial Allocation[d]	2017	Vector	(OpenStreetMap Contributers, 2017)
Distance to OpenStreet Map (OSM) Roads	osmroa_dst	Distance to nearest OSM road feature	Spatial Allocation[d]	2017	Vector	(OpenStreetMap Contributers, 2017)
Average Precipitation	wclin_prec	Mean Precipitation	Spatial Allocation[d]	1950 - 2000	30 arc sec	(Hijmans, Cameron, Parra, Jones, & Jarvis, 2005)
Average Temperature	wclim_temp	Mean temperature	Spatial Allocation[d]	1950 - 2000	30 arc sec	(Hijmans et al., 2005)

a Some classes were collapsed: 10-30 → 11; 40-120 → 40; 150-153 → 150; 160-180 → 160 (Sorichetta et al., 2015)

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
(see d) 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 utilised 2000, 2005, 2010, and 2015 as observed points in the dasymetric modelling process, but only derived covariates for 2000 were utilised in the random forest as predictive covariates

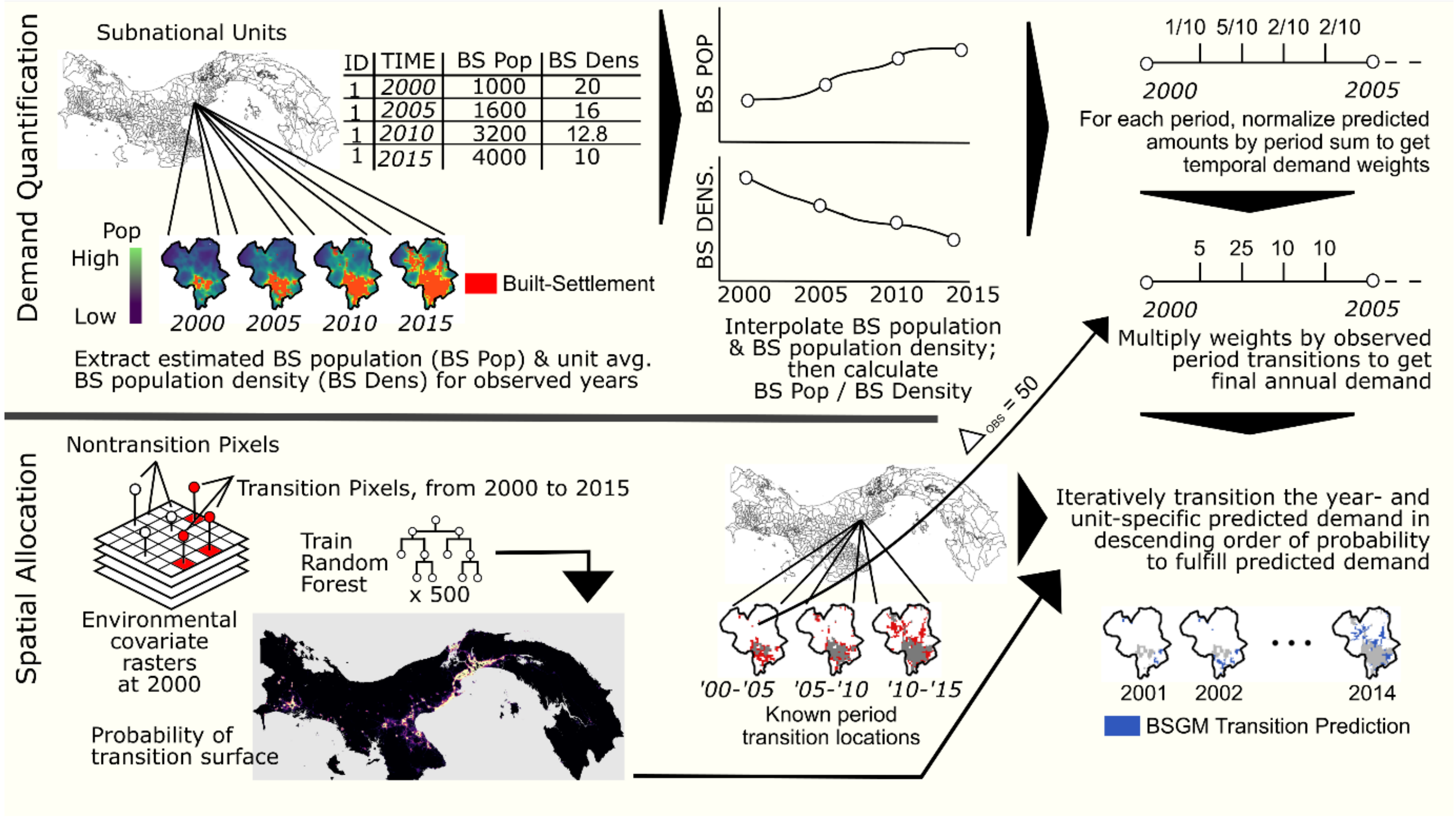
d Used as predictive covariates in the random forest calculated probabilities of transition

2.5 Built-Settlement Growth Model (BSGM)

2.5.1 Overview

Here we interpolated BS extents for every year between a set of RS-based observed years, *T* = {*t0*, t_1, t_2, *…*, *t1*} where *t0* is the initial RS-based observed year, *t1* is the final RS-based observed year, and all other times t_k are years lying between *t0* and *t1* for which we had RS-based observed BS extents. The time between any two RS-based observed time points *t* is referred to as a period, *p*, with all periods being a subset of *P*. Within this study, *T* = {2000, 2005, 2010, 2015} and *P*={2000-2005, 2005-2010, 2010-2015} and, therefore, we are modelling across three periods, estimating BS extents for 12 years, based upon the input of four RS-based observed years. However, the interpolative BSGM modelling framework can handle any regularly spaced intra-period time-step if the input data corresponds.

The interpolative BSGM modelling framework has two main components: a demand quantification component and a spatial allocation component, as shown in Figure 2.


Demand Quantification
Subnational Units
Pop
High
Low
2000 2005 2010 2015
Built-Settlement
Extract estimated BS population (BS Pop) & unit avg. BS population density (BS Dens) for observed years
BS POP
BS DENS.
2000 2005 2010 2015
Interpolate BS population & BS population density; then calculate BS Pop / BS Density
1/10 5/10 2/10 2/10
2000
2005
For each period, normalize predicted amounts by period sum to get temporal demand weights
5 25 10 10
2000
2005
Multiply weights by observed period transitions to get final annual demand
Δ OBS = 50
Spatial Allocation
Nontransition Pixels
Transition Pixels, from 2000 to 2015
Train Random Forest
x 500
Environmental covariate rasters at 2000
Probability of transition surface
'00-'05 '05-'10 '10-'15
Known period transition locations
Iteratively transition the year- and unit-specific predicted demand in descending order of probability to fulfill predicted demand
2001 2002 ... 2014
BSGM Transition Prediction

ID	TIME	BS Pop	BS Dens
1	*2000*	1000	20
1	*2005*	1600	16
1	*2010*	3200	12.8
1	*2015*	4000	10

Figure 2. High-level example overview of the BSGM modelling framework process for interpolation using four RS-based observed years (2000, 2005, 2010, 2015) and predicting for all unobserved years in between. Note, example maps and numbers are not to scale.

We generalise the process to determine the number of non-BS to BS transitions for each year we are interpolating, i.e. demand quantification, independently for each subnational unit, hereafter unit, as follows:

1. Create a population map for all years in *T* (2000, 2005, 2010, 2015).
2. At all RS-based observed years *t* in *T*, for each unit, extract the time- and unit-specific population count within the corresponding BS extents and derive the corresponding unit-average BS population density, i.e. lacking more precise information all BS pixels in a unit have the same BS population density (Figure 2).
3. On a unit-by-unit basis, interpolate the extracted BS population count and BS population density for all years, t_k, between each RS-based observed year *t* in *T* (Figure 2).
4. Estimate year- and unit-specific number of expected non-BS-to-BS transitions based upon the corresponding predicted BS population and BS population density (Figure 2).
5. Within each unit, for each period, create annual demand weights by normalising the annual number of expected transitions (from step 4) by the sum of the period's annual number of expected transitions (Figure 2).
6. For each unit and period, use the annual weights (from step five) to dasymetrically redistribute the period's total observed transitions to each year within the given period (Figure 2). Repeat for all periods.

To spatially allocate, i.e. disaggregate, the estimated annual transitions, (from step 5) we first train a Random Forest (RF) model (Breiman, 2001) to produce a continuous surface representing the probability of a given pixel transitioning from non-BS to BS between *t0* and *t1*, i.e. 2000 and 2015 (Figure 2). For every year, and independently for each unit, we utilised unit-normalised annually lagged lights-at-night (LAN) data to adjust the base RF-derived transition probabilities annually. Given that the BSGM modelling framework is interpolative, we limited the spatial allocation component to predicting transition probabilities in pixels that, based upon the input data, were observed to have transitioned within the given period (Figure 2). For example, only pixels seen to have transitioned between 2000 and 2005 could be predicted as transitioning in 2001, 2002, 2003, or 2004. With this in mind, within each unit, we selected pixels with the *n*th highest probabilities for transition, where *n* was equal to the number of pixels estimated to transition in that unit for that year. 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 annual binary BS extent datasets. All modelling and analyses were carried out using R 3.4.2 (R Core Team, 2016) on the IRIDIS 4 high-performance computing cluster (see Appendices for the full process diagram and the Supplemental Materials for the modelling code).

2.5.2 Demand Quantification

First, we created population distribution datasets for all years *t* in T by using available time-specific covariates (see Appendices) and the method, described in Gaughan et al. (2016) and Stevens et al. (2015), to dasymetrically redistribute the time-specific unit-based population counts to 3 arc second grid pixels (Mennis, 2003; Mennis & Hultgren, 2006). Second, for each unit and year *t* in *T*, we extracted and summed the population counts spatially coincident with the BS extents, i.e. BS population counts, and derived the corresponding BS population density for use in the later stages of the demand quantification component. Third, for each year t_p within a given period *p*, we interpolated the BS population count of each unit *i*, i.e. $BSPOP_i(t_p)$, using logistic growth curves with year-specific total population, $K_i(t_p)$, as the dynamic carrying capacity (Booth, 2006; P. S. Meyer & Ausubel, 1999). See Appendices for rationale regarding the use of a logistic growth curve with a dynamic limiting factor. These curves were fitted in a piecewise manner, i.e. one curve for each period *p* ∈ *P*. This is written in Equation 1 as:

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

where r_i and C_i are determined by fitting a least-squares linear regression to the set of observed values corresponding the given period after having been transformed via Equation 2:

$$\ln\left(\frac{BSPOP_{it_{observed}}}{K_{it_{observed}} - BSPOP_{it_{observed}}}\right) = r_i(t_p) + C_i \qquad \text{[Eq. 2]}$$

Fourth, to interpolate the unit-average BS population density for each unobserved year t_k between the years *t* in *T*, we fit natural cubic splines (McNeil, Trussell, & Turner, 1977) for each unit *i* across all unobserved years using the years *t* in *T* as the knots. Our priority being that the fit curve would match our values of observation, adapting to the data, i.e. non-parametric smoothing, rather than adapting the data to a specific distribution, i.e. parametric approach. See Appendices for more rationale on the use of cubic splines.

Finally, to begin estimating number of transitions, in each unobserved year t_k and for each unit *i*, we simply related the corresponding interpolated BS population and BS population density in Equation 3:

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

where $BSD_i(t)$ is the unit-average BS population density at time *t*. See Appendices for how predicted “negative growth” resulting from Equations 1-3 was handled.

In order to maintain agreement with the input data, i.e. the RS-based observed BS extents, the sum of our annual estimated transitions needed to match the total number of observed transitions within a given period *p*. So, we reweighted the estimated transitions of each year on a unit-by-unit basis using the sum of the estimated transitions in the period *p*. To calculate the unit-and year-specific weight, $w_{ip}(t_p)$, within the period *p*, we write the calculation in Equation 4 as:

$$w_{ip}(t_p) \ = \frac{B\widehat{SCNT}_{\imath}(t_p)}{\sum_1^k B\widehat{SCNT}_{\imath}(t_p)} \qquad \text{[Eq. 4]}$$

where t_p is again relative to the given period *p*, from 1 to the last year *k*, and all w_{ip} for a given unit *i* and period *p* sum to one.

Then, using these weights, we carried out a temporal dasymetric redistribution of the total observed transitions from the larger source period *p*, e.g. 2000-2005, to the individual unobserved years, e.g. 2001, …, 2004. To obtain the final temporally disaggregated transitions, $BSCNT_{\imath_{FINAL}}(t_p)$, we multiplied the unit- and year-specific weight, $w_{ip}(t_p)$, by the corresponding period *p*'s observed transitions, $\Delta BSCNT_{ip}$, rounding to the nearest whole number for each year, as shown in Equation 5 (see Appendices for obtaining agreement with rounding differences).

$$BSCN\widehat{T_{\imath_{FINAL}}}(t_p) = round(w_{ip}(t_p) * \Delta BSCNT_{ip}) \qquad \text{[Eq. 5]}$$

2.2.2 Spatial Allocation

We utilised a RF model to accurately and efficiently model, across each country, the probability of each pixel transitioning from non-BS-to-BS. Importance of individual covariates in a classification random forest are typically measured by the average decrease in the Gini impurity, the probability of incorrectly classifying a random selected element of the dataset if it were randomly assigned label based upon the distribution of classes in the dataset (Breiman, 2001). A RF model was selected for its robustness to noise, its automatability and efficiency, and its ability to capture non-linear and complex interactions c. Furthermore, Kamusoko & Gamba (2015) showed that RFs have been shown to perform equally to, if not better, than other methods, (including support vector machines and logistic regression) used for predicting the probability of transitioning from non-built-environment to built-environment.

The binary dataset of non-BS-to-BS transition constitutes an intrinsic "imbalanced set" (He & Garcia, 2009), i.e. there are many more non-transitions than transitions. So, we adopted a stratified random over/under-sampling method (He & Garcia, 2009), similar to (Linard et al., 2013), as follows: (i) randomly sample 80 percent of the pixels observed to have transitioned between 2000 and 2015, up to 50,000 and, (ii) randomly sample an equal number of pixels that have not transitioned during the same time span. We then used these training sets and spatially and temporally coincident covariates to train a RF model for each country and predicted the corresponding surface of non-BS-to-BS transition probabilities. All covariates used were retained in the final model. These probabilities have a value between 0 and 1 and represent the posterior probability of a pixel being classified by the RF model as transitioning between *t0*, 2000, and *t1*, 2015 *c*.

We then refined these probabilities to annual probabilities using annual ancillary information. Given that changes in LAN brightness have been found to be good indicators of population and urban growth (Zhang & Seto, 2011), we adjusted the RF-derived transition probabilities using annual weights based upon unit-normalised annual average LAN brightness differences prior to spatially disaggregating the estimated annual non-BS-to-BS transitions from the demand quantification component. The rationale being that larger increases in average annual brightness for a given pixel, relative to all other pixels within the same unit, represent a higher relative probability of non-BS-to-BS transitions for that pixel and vice versa.

To create these annual spatial weights, we first calculated the annual lags of the LAN radiance values and rescaled the differences between 0 (unit's lowest value) and 1 (unit's highest value). This rescaling was based upon the values of all pixels *M* within a given unit *i* for a given lag *l*, where the number of lags is equal to the number of years minus one, e.g. for 2000 to 2015 we have 14 lags beginning with 2001 minus 2000. This calculation for a given pixel *m*, where *m*∈*M* pixels total in the unit, can be written as:

$$wLAN_{i,m,l} = \frac{lag_{i,m,l} - \min\left(lag_{i,M,l}\right)}{\max\left(lag_{i,M,l}\right) - \min\left(lag_{i,M,l}\right)} \qquad \text{[Eq. 6]}$$

where $lag_{m,l} = LAN_{m,\tau} - LAN_{m,\tau-1}$ and τ represents the most recent year of the lag *l*, e.g. for lag 2001-2000 τ would be 2001. We then calculated year specific transition probabilities for every pixel known to have transitioned, *j*, using Equation 7:

$$P_{adj}(transition)_{ijt} = wLAN_{ijt} * P(transition)_{ij} \qquad \text{[Eq. 7]}$$

where $P(transition)_{ij}$ is the RF-derived transition probability for observed transition pixel *j* in unit *i* and $P_{adj}(transition)_{ijt}$ is the corresponding resultant adjusted transition probability for year *t*:

Using these adjusted probabilities, we then spatially disaggregated the estimated annual transitions, from the demand quantification component, within each unit. Given that the non-BS-to-BS transition process is iterative in nature, we began by taking the extents of the previous year. Within each unit *i* and for each period *p*, we limited the location(s) where transitions could be allocated to pixels *j* as defined by the RS-based observed BS extents. For all pixels *j*, assuming they were not transitioned in previously iterated years, we retrieved the adjusted transition probabilities and, similar to previous models (Linard et al., 2013; Tayyebi et al., 2013), we assumed pixels with a higher probability of transition were more likely to transition before pixels with lower probabilities. We selected the *n*th highest probabilities from the pixels *J* in unit *i*, where *n* was equal to $\widehat{BSCNT_{\iota}}_{FINAL}$, changed the value of those *n* pixels to represent a non-BS-to-BS transition, and output the union of the new transitions and previous BS extents as the predicted BS extents for that year. We repeated this procedure using the newly produced extents for the preceding year as the base BS extent for the next year's transition procedure, until all years for the given period *p* were processed and then the entire procedure was repeated until all periods *p* in *P* had been processed, resulting in annual modelled BS extents.

2.6 Analyses

2.6.1 Validation and Comparison Metrics

While the RF produces its own validation estimates (Breiman, 2001), we tested the accuracy of the RF classifier by randomly sampling 100,000 pixels, not utilised 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 (He & Garcia, 2009; Saito & Rehmsmeier, 2015), Precision Recall Curves (PRCs) with simulated perfect and random classifier curves for comparison.

Here we validated the modelled BS extents to all withheld ESA RS-based BS extents corresponding to the unobserved years between 2000 and 2015, i.e. 2001-2004, 2006-2009, and 2011-2014. Here “True” represents agreement of the

BSGM-based BS extents to the temporally corresponding withheld annual ESA RS-based BS extents and vice versa. For every year of prediction, we determined whether a pixel was True Positive (TP), False Positive (FP), False Negative (FN), or True Negative (TN). Pixels used for validation of the modelled BS extents were limited only to pixels observed transitioning from non-BS to BS between the modelled periods for two related reasons:

1) Being an interpolative model, we constrained the areas of possible transition 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 for any specific year between those periods remained.
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.

We calculated contingency table-based metrics to evaluate classification agreement based primarily on the F_1 score (Table 3) which is the harmonic mean of recall and precision, the quantity disagreement (R.G. Pontius & Millones, 2011), and the allocation disagreement (R.G. Pontius & Millones, 2011). We aggregated the pixel level results (See Supplemental Materials), 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) (Openshaw, 1984).The MAUP not only reduces variance in value distributions the more the data are aggregated from their original resolution (Openshaw, 1984), but will result in different prevalences within different units, i.e. zonal, configurations. The equations of the metrics calculated are listed in Table 3.

Table 3. Classification agreement metrics. 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) (Rogan & Gladen, 1978)	$\frac{TP}{TP+FN}$	0 (no recall) – 1 (perfect recall)
Specificity (Rogan & Gladen, 1978)	$\frac{TN}{FP+TN}$	0 (no specificity) – 1 (perfect specificity)
Quantity Disagreement (R.G. Pontius & Millones, 2011)	$\frac{\left\lvert\frac{FN-FP}{TP+FP+FN+TN}\right\rvert+\left\lvert\frac{FP-FN}{TP+FP+FN+TN}\right\rvert}{2}$	0 (no disagreement) – 1 (complete disagreement)
Allocation Disagreement (R.G. Pontius & Millones, 2011)	$2*\min\left(\frac{FP}{TP+FP+FN+TN},\frac{FN}{TP+FP+FN+TN}\right)$	0 (no disagreement) – 1 (complete disagreement)
F_1 score	$\frac{2*\frac{TP}{TP+FP}*\frac{TP}{TP+FN}}{\frac{TP}{TP+FP}+\frac{TP}{TP+FN}}$	0 (worst) – 1 (best)

As suggested by Pontius, Shusas, and McEachern (2004), to assess the predictive ability of the BSGM modelling framework, we compared it to 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 framework. Again, we determined whether each pixel was a TP, FP, FN, or TN and calculated metrics to compare the BSGM-based BS extents and the BS extents produced using the naive model for 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-based BS extents 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 one calculated for the BS extents produced using the naive model, but had more variance in absolute terms of performance. Comparable results between were found at the unit level (See Appendices), with relatively higher performance in the middle and later years of the study period.

3.1 RF Performance

The ROC plots (left plots in Figure 3) show that the RFs approach the performance of the theoretical perfect model. However, given the imbalanced data, the PRC plots (right plots in Figure 3) 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 quickly decreases with increasing recall.

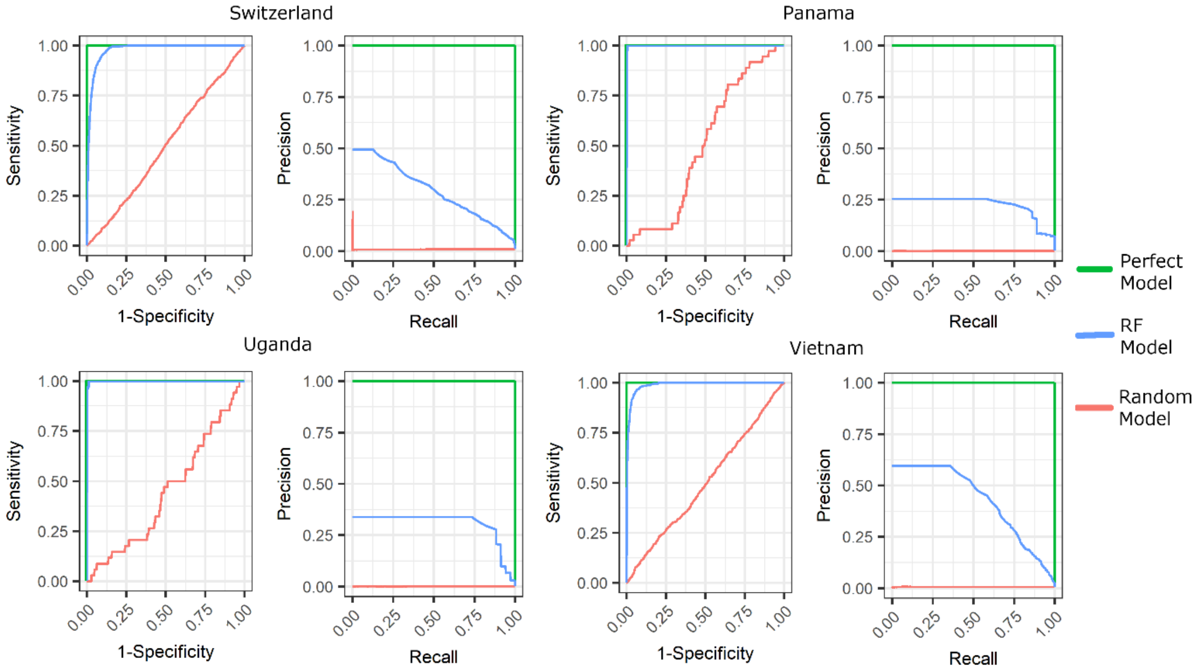


Figure 3. Receiver Operator Curve (left plots) and Precision Recall Curves (right plots) with the RF model performance, blue lines, against a random model (red lines), and a perfect model (green lines), for each modelled country.

Of the covariates informing the RF models, we consistently saw that the most important predictors of a pixel transitioning from non-BS to BS (Figure 4) were covariates related to distance (“esa_cls190_dst_2000”) and local density of BS (“esa_cls190_prp_5_2000”, “esa_cls190_prp_10_2000”, and “esa_cls190_prp_15_2000”) established at the beginning of the overall study period, i.e. 2000. Other important predictors included connectivity of BS extents ("tt50k_2000") at the beginning or approximately end ("urbanaccessibility_2015" and "osmroa_dst") of the study period (Figure 4).

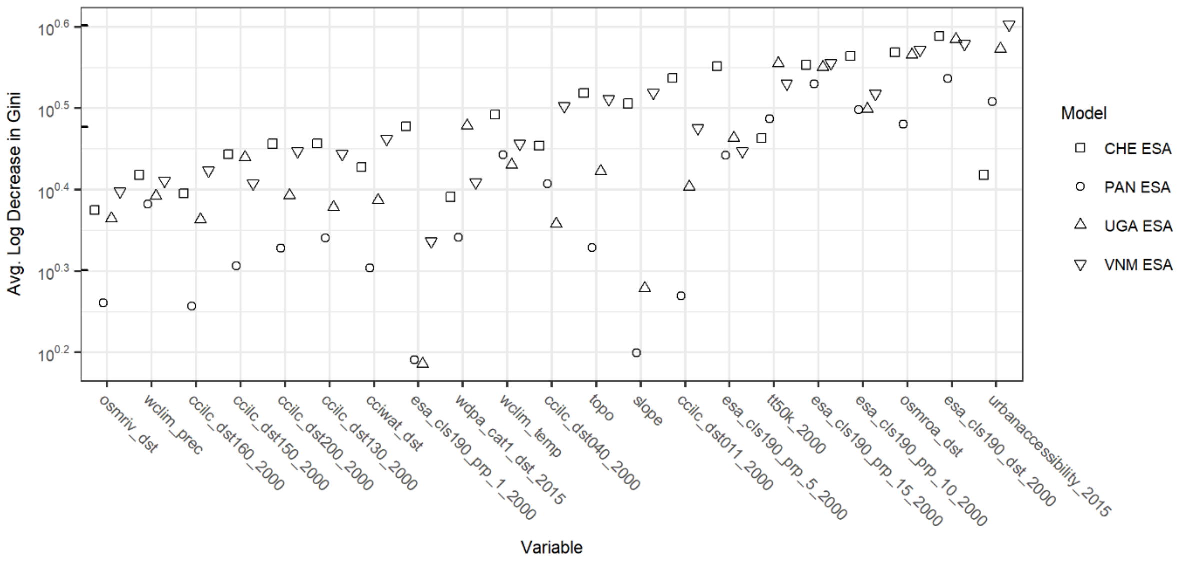


Figure 4. Random forest covariate importance as measured by the average log decrease in the Gini impurity when the covariate is used as the splitting criteria at nodes, for Swizerland (CHE) ESA, Panama (PAN) ESA, Uganda (UGA) ESA, and Vietnam (VNM). Higher values indicate better predictive performance of covariate. Refer to Table 2 for covariate names.

3.2 Predicted BS Extents Results

Examining the proportion of pixels known to transition that were predicted correctly (Table 4), we show that out of 48 modelled BS extents (corresponding to 12 years across four countries), 39 of those had correctly predicted proportions between 0.80 and 0.99 (green) with 25 of them having proportions over 0.90. Modelled extents 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 (R.G. Pontius & Millones, 2011).

Table 4. Proportion of transition pixels predicted correctly by the BSGM modelling framework by year for Switzerland (CHE, Panama (PAN), Uganda (UGA), and Vietnam (VNM). Modelled extents 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 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

Further examining source of disagreement, we display the quantity and allocation disagreement between the BSGM-based RS extents and validation set, i.e. ESA RS-based BS extents, as well as the corresponding disagreements with the BS extents produced using the naïve model (Figure 5). We show that for all modelled years the total disagreement is substantially less than that of the naive model and the disagreement produced by the BSGM modelling framework is predominantly due to quantity error (Figure 5). However, there does appear to be a pattern of increasing disagreement due to allocation error after 2010. Identical analyses for the early WSF Evolution data are provided in Appendices.

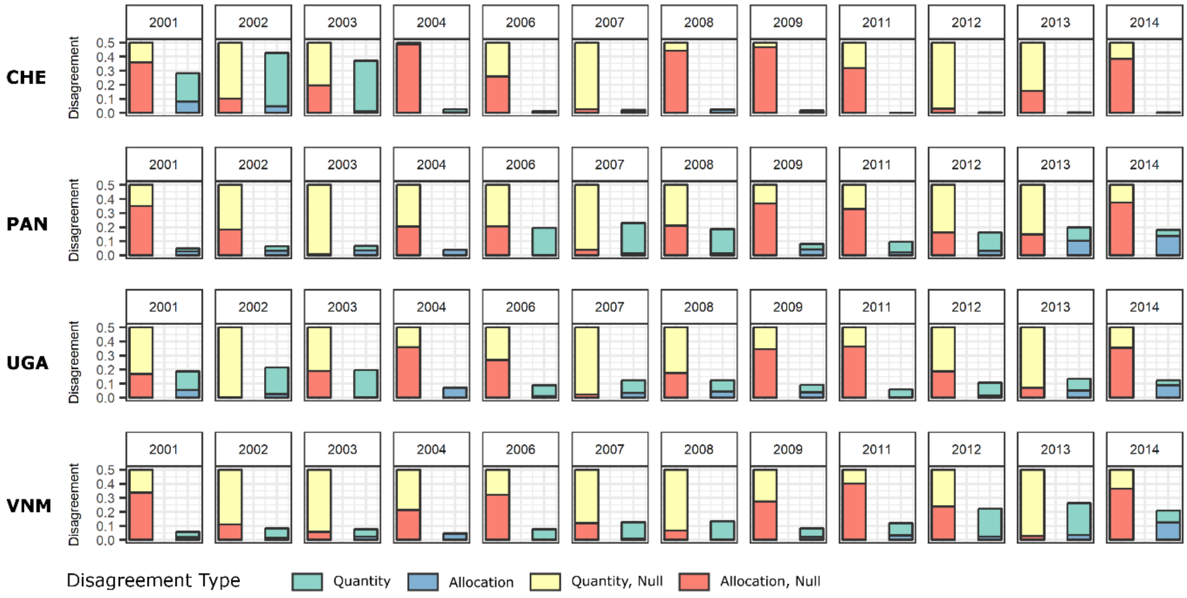


Figure 5. Pixel-level quantity and allocation disagreement of BSGM and naive models for Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) as compared to a naive model, given in yellow and red. Full annual contingency data and metrics in supplemental materials.

While our ESA RS-based BS extents data does not give information on the true size of settlements on the ground, we can leverage the fact that our subnational units are derived from census boundaries (Doxsey-Whitfield et al., 2015), which are known to typically be smaller in areas of larger settlements and larger in areas of more fragmented smaller settlements, to begin understanding how the framework is operating across the continuum of settlement size. Looking at the contour density plots of F1 unit-level scores across all years for each country plotted against the corresponding subnational unit area, in Figure **6**, we can see that higher scores are clustered for units with smaller areas across each country, although, with the exception of Uganda, the framework shows good density and performance over a range of unit sizes. Less variance in performance for larger units is likely due to the smaller amount of transitions seen in these units, decreasing the probabilities for error by the interpolative BSGM modelling framework.

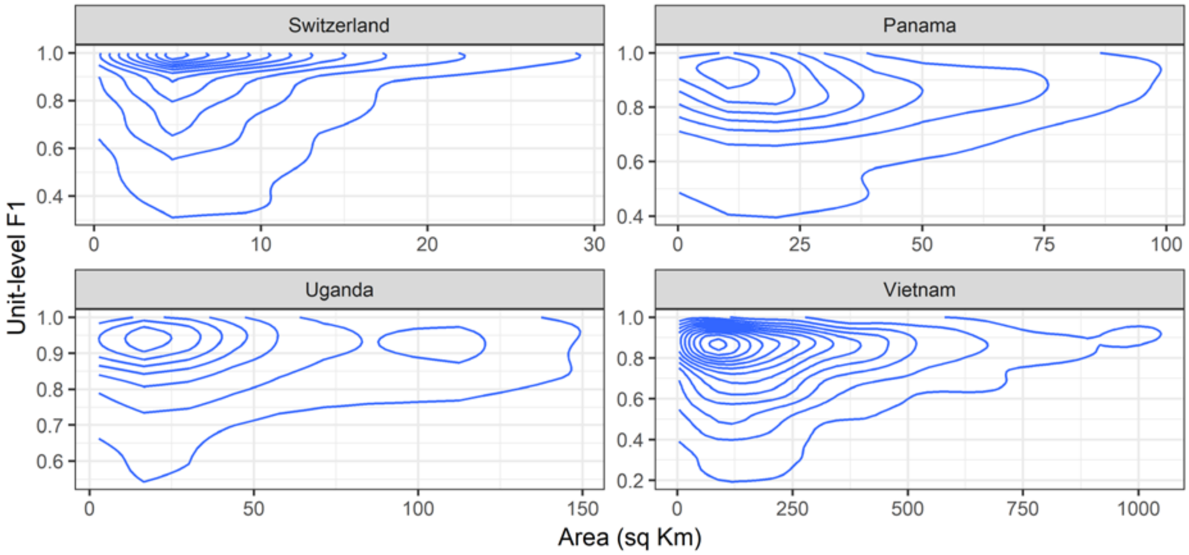


Figure 6. Contour density plot of unit-level F1 scores by country across all predicted years. Created using a two-dimensional kernel density estimation (Venables & Ripley, 2002).

Examining examples of the annual BSGM-based BS extents against the corresponding annual ESA RS-based BS extents for the mid-point of each period, which should theoretically be the worst simply by being the furthest year from any observation we note a few things of interest. First, there are relatively large amounts of agreement whether for small or large settlements (with Visp being a town of less than 10,000). Second, the framework seems to predict “infill” growth, e.g. Kampala in 2003 and North Ho Chi Minh City in 2013, later than indicated by the corresponding ESA RS-based extents (Figure 7, in red). Lastly, it appears that the BSGM modelling framework is temporally conservative in that it is not predicting relatively large amounts of pixels too early (Figure 7, in blue). Of course, the model performance can vary from unit-to-unit and year-to-year, and we provide the entire annual BSGM-based BS extents in GeoTiff format in the Supplemental Materials.

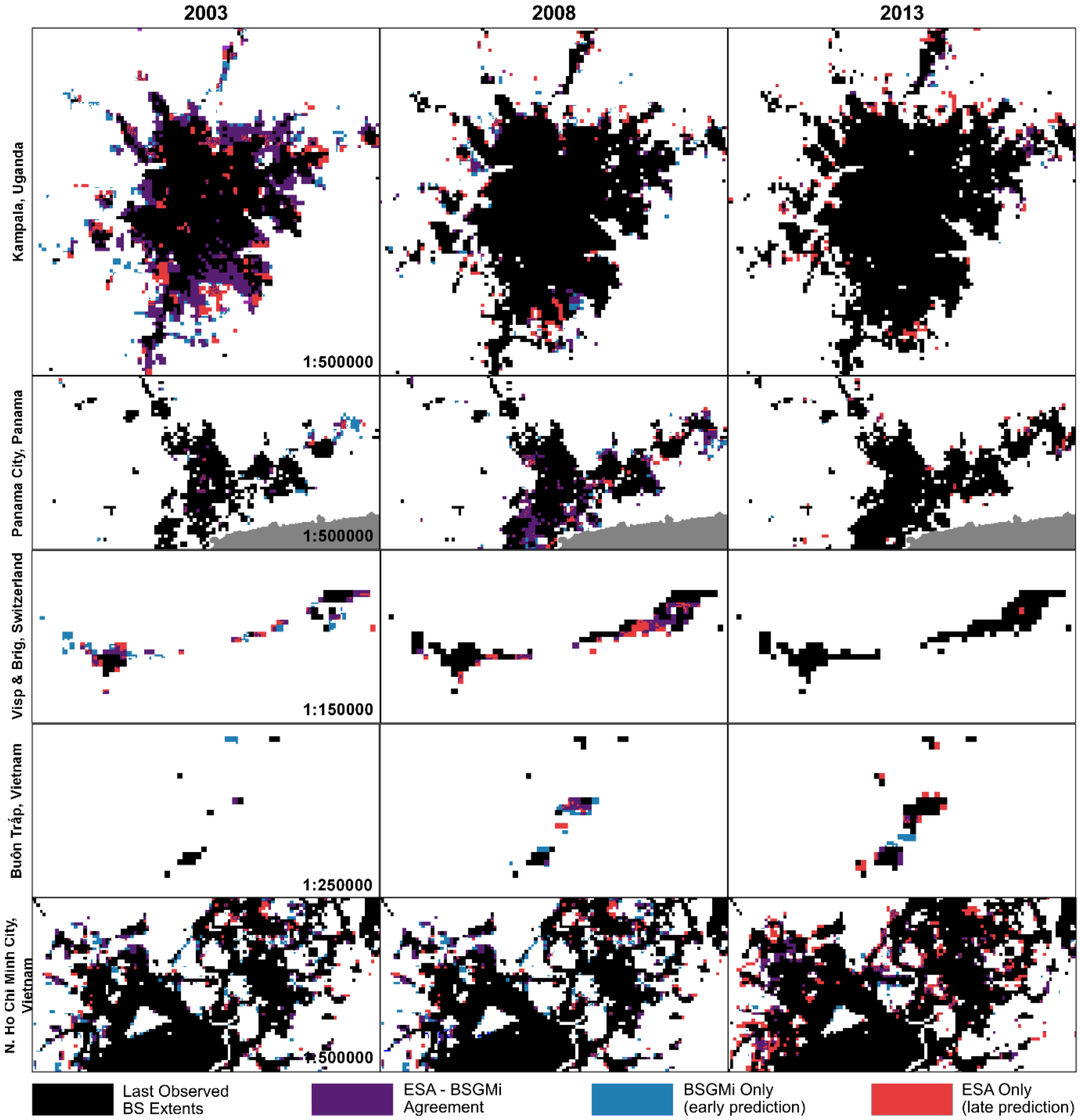


Figure 7. Selected BSGM-based BS extent and ESA RS-based BS-extent used for validation across the four countries for the approximate mid-point years of each period – 2003, 2008, 2013. "ESA Only" represents BS pixels in the validation dataset not classified as BS pixels in the corresponding BSGM-based BS extent. "BSGMi Only" represents BS pixels in the BSGM-based BS extent not classified as BS pixels in the validation dataset.

4. DISCUSSION

Here we have shown that the BSGM framework is capable of filling gaps in time-series of built-settlement datasets by estimating the extents in between RS-based imagery using relative changes in BS population and BS population density combined with environmental covariates. The BSGM modelling framework approximates patterns of BS growth through time with good agreement to its input BS extent dataset for most years, both at the pixel and unit level (Table 4, Figures 5 -7, and Appendices). This emphasizes the strength of incorporating the use of an interpolative model, such as the BSGM modelling framework, as opposed to solely using urban feature datasets that are largely imagery-dependent. While still the gold standard, these imagery-based datasets may be affected by 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 population distribution through time, inform future extractions of BS 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, this validation design has limits. The agreements and disagreements here are generated by how well the BSGM model replicates the spatio-temporal data patterns of the input ESA BS extents and does not state anything about the accuracy of the BSGM-predicted extents as compared to ground truth. Even if we possessed accurate and time-specific BS ground truth extents with complete spatial coverage, given that the BSGM is an interpolative modelling framework it would be difficult to determine if any error originated from the model or was propagated from the input BS extents. Performance as assessed by ground truth would be highly sensitive to the chosen BS extents input to the BSGM. We assume if the BSGM can accurately replicate and interpolate the data patterns of the input dataset, then the end user can have some confidence that validation metrics provided by the original data producers, e.g. ESA, are likely to hold. However, ground truth accuracy is important for some end users, and we encourage them to assess the BSGM output accordingly where data allows.

The BSGM is neither without error nor a replacement for urban feature/built-settlement extractions methods. Given that the BSGM modelling framework is interpolative, its modelled BS extents are limited by the accuracy, the spatial and temporal resolution of its inputs including the RS-based observed BS extents, the time specific subnational population data, and the spatially-explicit population distribution dataset. For example, the poorer model performance from 2001 through 2003 (Table 4 and Figures 5) is likely due to the fact the ESA RS-based BS extents were delineated at 30 arc sec resolution, due to the MERIS and PROBA V imagery not being available, rather than the 10 arc second resolution for years from 2004 through 2015 (UCL Geomatics, 2017). With regards to the total disagreement of the BSGM-based BS extents to the ESA RS-based BS extents (Figure 5), the relatively low contribution of allocation disagreement prior to circa 2010 and corresponding increase in contribution post-2010 is possibly due to the switch from using coarser DMSP-based LAN data to VIIRS-based LAN data at the 2012 time point.

The BSGM modelling framework is also limited by conceptual and mathematical assumptions. We are assuming a certain relationship between relative BS population and BS population density changes and drive demand for temporally coincident BS growth. Furthermore, we assume that BS population grows logistically with a time varying capacity that is temporally coincident and that 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 BS growth (Angel et al., 2011; Dyson, 2011; Linard et al., 2013; Seto et al., 2011, 2012), there is also evidence for considering other drivers, not used here because of their unavailability at subnational levels globally through time, such as Gross Domestic Product and arable land per capita (Angel et al., 2011; Seto et al., 2011). Furthermore, there are other “intangibles” such as local, regional, and national land use or development policies, which almost certainly shape the BS growth, but are typically not available in an accessible format or not available at all. Furthermore, the BSGM modelling framework is relying on temporally (Doxsey-Whitfield et al., 2015) and spatially (Stevens et al., 2015) modelled subnational population data that are used as inputs to estimate the BS population at each point in time. However, regardless of the modelling approach used to spatially disaggregate the population from the unit to the pixel-level, since the BSGM modelling framework allocates transitions based upon relative changes in BS population, the errors associated with the spatial redistribution of the population should not affect prediction timings, as long as biases are consistent over times. As with any “model outputs built upon model outputs,” users of such datasets must be cautious of accumulated errors.

When considering area-based metrics, the Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984) must be considered. Indeed, the total number of pixels in each unit is typically larger in the less settled units, resulting in less variation of aggregated metric values referring to those. With dasymetric redistribution methods, the size and spatial arrangement of the source units, can also affect the quality of the disaggregation with the larger relative differences between source unit and target unit sizes introducing more (Mennis, 2003; Mennis & Hultgren, 2006). This, in part, likely led to the results in Figure 6. The MAUP could also explain the framework's late prediction regarding infill growth (Figure 7), with the unit-averaging of the BS population density potentially obscuring the underlying sub-unit variation (Openshaw, 1984) in BS population density that could be more likely driving pixel-level non-BS-to-BS transitions. Other reasons for disagreements in Figure 7 could be due to less detectable light changes associated with non-BS-to-BS transition due to light blooming. However, annual BSGM-based BS extents can be aggregated across years to decrease uncertainty of the interpolated extents, as the growth of BS extents is an incremental process, with future outcome dependent upon previous growth.

Unfortunately, for both the use of logistic curves to interpolate between estimates of BS population count data and the use of cubic splines to interpolate between estimates of BS population density data, independent data does not exist to evaluate the error or uncertainty of the interpolated values. 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 RS-based observed transition counts across time (Equations 1-7). Finally, it is important to highlight how this dasymetric disaggregation by weights precludes the propagation of any uncertainties calculated before the disaggregation step (as 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

The 2030 Agenda for Sustainable Development and its SDGs have reinforced the importance of data to being able to account for "all people everywhere (United Nations, 2016). Differences in the dynamic spatial distributions of hazards (Carrão, Naumann, & Barbosa, 2016; Oliveira, Oehler, San-Miguel-Ayanz, Camia, & Pereira, 2012), the spatial variation of the effects of climate change (Ericson, Vorosmarty, Dingman, Ward, & Meybeck, 2006; Hanjra & Qureshi, 2010; Stephenson et al., 2010), spatially allocating services to ensure sufficient coverage (Eckert & Kohler, 2014; Sverdlik, 2011), and targeting interventions and planning (Linard, Alegana, Noor, Snow, & Tatem, 2010; Utazi et al., 2018) based upon local context with limited resources requires higher temporal resolution in the mapping of BS and mapping of populations, both large and small (United Nations, 2016). Here we described a flexible modelling framework for globally modelling BS extents between RS-based observed time points, with 39 of 48 validated BSGM predicted BS extents having over 80 percent agreement with ESA RS-based observed extents and 25 of those years having over 90 percent agreement (Table 4). This framework is scalable globally, but also allows for sub-national variation in transition probability, population changes, and local relative LAN changes to drive the overall study area model.

As global urban feature/built-settlement extent datasets such as ESA CCI, MAUPP, GHSL, GUF and others continue to improve both in terms of spatial accuracy and spatial and temporal resolutions, modelling frameworks such as the BSGM will likely still be useful due to imagery/extraction issues and the need to smooth or fill-in time-series of urban feature/built-settlement datasets (ESA CCI, 2017; Esch et al., 2013; Forget, Linard, & Gilbert, 2018; Pesaresi et al., 2016). By the time annual urban feature/built-settlement extractions from currently available imagery will become an economically viable means of filling gaps, the current demand for annual datasets, eventually becoming the standard, will be replaced by grown a demand for quarterly and monthly datasets. This is not to say that interpolative models and feature-extraction algorithms are oppositional, but rather that they are complementary. Should the time come where high-resolution global annual urban feature/built-settlement datasets become the norm, this would offer a wealth of information from which to improve the assumptions the BSGM currently makes.

As informative as global RS-based urban feature/built-settlement datasets are, imagery will never see into the future and we plan on extending the BSGM modelling framework to allow for short-term projection of the growth of BS extents. 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 4) giving support to work in attempting to define “urban” based on contiguity, connectivity, and spatial density (Dijkstra & Poelman, 2014; Esch et al., 2014; Pesaresi & Freire, 2016). Still mostly unknown is how the BSGM modelling framework would perform for smaller settlements, not captured by the coarser datasets such as the ESA CCI land cover, and we are looking to test this with forthcoming feature data sets with resolutions below 3 arc seconds. Further sensitivity testing of the framework to noisy or biased inputs, e.g. BS datasets in arid biomes, is also planned. Lastly, we plan to validate the utility of these dataset in an applied manner by comparing the effects of including the BSGM-based BS extents in annual population distribution modelling. Finally, the BSGM modelling framework can be adapted to run at other scales, both spatially and temporally, either by modifying the provided code (See Supplemental Materials) 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 model code and results datasets used here provided in the Supplemental Material.

REFERENCES

Angel, S., Parent, J., Civco, D. L., Blei, A. M., & Potere, D. (2011). The Dimensions of Global Urban Expansion: Estimates and Projections for All Countries, 2000-2050. *Progress in Planning*, *75*, 53–107. https://doi.org/10.1016/j.progress.2011.04.001

Angel, S., Sheppard, S. C., & Civco, D. L. (2005). *The Dynamics of Global Urban Expansion*. Washington, D. C.: The World Bank.

Barredo, J. I., Demicheli, L., Lavalle, C., Kasanko, M., & McCormick, N. (2004). Modelling Future Urban Scenarios in Developing Countries: An Application Case Study in Lagos, Nigeria. *Environment and Planning B*, *31*, 65–84.

Bartholomé, E., & Belward, A. S. (2005). GLC2000: a new approach to global land cover mapping from Earth observation data. *International Journal of Remote Sensing*, *26*(9), 1959–1977. https://doi.org/10.1080/01431160412331291297

Batty, M. (2009). Urban Modeling. In *International Encyclopedia of Human Geography* (pp. 51–58). Oxford, UK: Elsevier.

Batty, M., & Xie, Y. (1994). From Cells to Cities. *Environment and Planning B*, *21*, S31–S48.

Berechman, J., & Gordon, P. (1986). Linked Models of Land-Use Transport Interactions: A Review. In B. Hutchinson & M. Batty (Eds.), *Advances in Urban Systems Modelling*. Elsevier Ltd.

Booth, H. (2006). Demographic forecasting: 1980 to 2005 in review. *International Journal of Forecasting*, *22*(3), 547–581. https://doi.org/10.1016/j.ijforecast.2006.04.001

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Burgess, E. W. (1925). *The Growth of a City: An Introduction to a Research Project*. Chicago, IL: University of Chicago Press. https://doi.org/10.1080/00343404200021114

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. *Global Environmental Change*, *39*, 108–124. https://doi.org/10.1016/j.gloenvcha.2016.04.012

Chongsuvivatwong, V., Phua, K. H., Yap, M. T., Pocock, N. S., Hashim, J. H., Chhem, R., … Lopez, A. D. (2011). Health and health-care systems in southeast Asia: diversity and transitions. *The Lancet*, *377*(9763), 429–437. https://doi.org/10.1016/S0140-6736(10)61507-3

Clarke, K. C., & Gaydos, L. (1998). Loose-coupling a Cellular Automaton Model and GIS: Long-term Urban Growth Prediction for San Francisco and Washington/Baltimore. *International Journal of Geographic Information Sciences*, *12*(7), 699–714.

Clarke, K. C., Hoppen, S., & Gaydos, L. (1997). A Self-modifying Cellular Automaton Model of Historical Urbanisation in the San Francisco Bay Area. *Environment and Planning B*, *24*, 247–261.

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

Cohen, B. (2006). Urbanisation in Developing Countries: Current Trends, Future Projections, and Key Challenges for Sustainability. *Technology in Society*, *28*, 63–80.

Dhingra, M. S., Artois, J., Robinson, T. P., Linard, C., Chaiban, C., Xenarios, I., … Gilbert, M. (2016). Global mapping of highly pathogenic avian influenza H5N1 and H5Nx clade 2.3.4.4 viruses with spatial cross-validation. *ELife*, *5*. https://doi.org/10.7554/eLife.19571

Dijkstra, L., & Poelman, H. (2014). *A harmonized definition of cities and rural areas: the new degree of urbanisation* (Regional Working Paper No. WP 01/2014). Retrieved from http://ec.europa.eu/regional_policy/sources/docgener/work/2014_01_new_urban.pdf

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O., & Baptista, S. R. (2015). Taking advantage of the improved availability of census data: A first look at the Gridded Population of the World, Version 4. *Papers in Applied Geography*, *1*(3), 226–234. https://doi.org/10.1080/23754931.2015.1014272

Dyson, T. (2011). The role of the demographic transition in the process of urbanisation. *Population and Development Review*, *37*(Supplement), 34–54.

Earth Observation Group NOAA National Geophysical Data Center. (2016). VIIRS Nighttime Lights - One Month Composites. Boulder, Colorado: NOAA National Centers for Environmental Information. Retrieved from https://ngdc.noaa.gov/eog/viirs/download_dnb_composites.html

Eckert, S., & Kohler, S. (2014). Urbanisation and Health in Developing Countries: A Systematic Review. *World Health & Population*, *15*(1), 7–20.

Elvidge, C. D., Baugh, K. E., Kihn, E. A., Kroehl, H. W., & Davis, E. R. (1997). Mapping city lights with nighttime data from the DMSP Operational Linescan System system. *Photogrammetric Engineering & Remote Sensing*, *63*(6), 727–734.

Ericson, J. P., Vorosmarty, C. J., Dingman, S. L., Ward, L. G., & Meybeck, M. (2006). Effective sea-level rise and deltas: Causes of change and human dimension implications. *Global and Planetary Change*, *50*, 63–82.

ESA CCI. (2017). European Space Agency Climate Change Initiative Landcover. European Space Agency. Retrieved from http://maps.elie.ucl.ac.be/CCI/viewer/download.php

Esch, T., Bachofer, F., Heldens, W., Hirner, A., Marconcini, M., Palacios-Lopez, D., … Gorelick, N. (2018). Where We Live—A Summary of the Achievements and Planned Evolution of the Global Urban Footprint. *Remote Sensing*, *10*(6), 895. https://doi.org/10.3390/rs10060895

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., … Dech, S. (2013). Urban Footprint Processor - Fully Automated Processing Chain Generating Settlement Masks from Global Data of the TanDEM-X Mission. *IEEE Geoscience and Remote Sensing Letters*, *10*(6), 1617–1621.

Esch, T., Marconcini, M., Marmanis, D., Zeidler, J., Elsayed, S., Metz, A., & Dech, S. (2014). Dimensioning urbanisation - An advanced procedure for characterizing human settlement properties using spatial network analysis. *Applied Geography*, *55*, 212–228. https://doi.org//j.apgeog.2014.09.009

Forget, Y., Linard, C., & Gilbert, M. (2018). Supervised Classification of Built-Up Areas in Sub-Saharan African Cities Using Landsat Imagery and OpenStreetMap. *Remote Sensing*, *10*(7), 1145. https://doi.org/10.3390/rs10071145

Gaughan, A. E., Stevens, F. R., Huang, Z., Nieves, J. J., Sorichetta, A., Lai, S., …

Tatem, A. J. (2016). Spatiotemporal patterns of population in mainland China, 1990 to 2010. *Scientific Data*, *3*. https://doi.org/10.1038/sdata.2016.5

Goldewijk, K. K., Beusen, A., & Janssen, P. (2010). Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1. *The Holocene*, *20*(4), 565–573. https://doi.org/10.1177/0959683609356587

Gottman, J. (1957). Megalopolis, or the urbanisation of the north eastern seaboard. *Economic Geography*, *33*, 189–200.

Haas, H. De. (2010). Migration and Development: A Theoretical Perspective. *International Migration Review*, *44*(1), 227–264. https://doi.org/10.1111/j.1747-7379.2009.00804.x

Hanjra, M. A., & Qureshi, M. E. (2010). Global Water Crisis and Future Food Security in an Era of Climate Change. *Food Policy*, *35*, 365–377. https://doi.org/10.1016/j.foodpol.2010.05.006

Harris, C. D., & Ullman, E. L. (1945). The nature of cities. *Annals of the American Academy of Political and Social Sciences*, *242*, 7–17.

He, H., & Garcia, E. A. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Henderson, M., Yeh, E. T., Gong, P., Elvidge, C. D., & Baugh, K. (2003). Validation of urban coundaries derived from global night-time satellite imagery. *International Journal of Remote Sensing*, *24*(3), 595–609. https://doi.org/10.1080/01431160304982

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, *25*, 1965–1978.

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

Huang, X, Wen, D., Li, J., & Qin, R. (2017). Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery. *Remote Sensing of Environment*, *196*, 56–75. https://doi.org/10.1016/j.rse.2017.05.001

Huang, Xin, Hu, T., Li, J., & Wang, Q. (2018). Mapping Urban Areas in China Using Multisource Data With a Novel Ensemble SVM Method. *IEEE Transactions on*

Geoscience and Remote Sensing, *56*(8), 4258–4273. https://doi.org/10.1109/TGRS.2018.2805829

Kamusoko, C., & Gamba, J. (2015). Simulating Urban Growth Using a Random Forest-Cellular Automata (RF-CA) Model. *ISPRS International Journal of Geo-Information*, *4*, 447–470.

Lamarche, C., Santoro, M., Bontemps, S., D'Andrimont, R., Radoux, J., Giustarini, L., … Arino, O. (2017). 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 Sensing*, *9*(36). https://doi.org/10.3390/rs9010036

Leao, S., Bishop, I., & Evans, D. (2004). Simulating Urban Growth in a Developing Nation's Region Using a Cellular Automata-based Model. *Journal of Urban Planning and Development*, *130*(3), 145–158.

Ledent, J. (1982). Rural-Urban Migration, Urbanisation, and Economic Development. *Economic Development and Cultural Change*, *30*(3), 507–538. Retrieved from https://www.jstor.org/stable/3203205

Lehner, B., Verdin, K., & Jarvis, A. (2008). New Global Hydrography Derived from Spaceborne Elevation Data. *Eos, Transactions of the American Geophysical Union*, *89*(10), 93–94. https://doi.org/10.1029/2008EO100001

Li, X., & Gong, P. (2016). Urban growth models: progress and perspective. *Science Bulletin*, *61*(21), 1637–1650. https://doi.org/10.1007/s11434-016-1111-1

Linard, C., Alegana, V., Noor, A. M., Snow, R. W., & Tatem, A. J. (2010). A high resolution spatial population database of somolia for disease risk mapping. *International Journal of Health Geographics*, *9*(1), 45.

Linard, C., Tatem, A. J., & Gilbert, M. (2013). Modelling Spatial Patterns of Urban Growth in Africa. *Applied Geography*, *44*, 23–32.

Liu, Xiaoping, Hu, G., Chen, Y., Li, X., Xu, X., Li, S., … Wang, S. (2018). High-resolution multi-temporal mapping of global urban land using Landsat images based on the Google Earth Engine Platform. *Remote Sensing of Environment*, *209*, 227–239. https://doi.org/10.1016/j.rse.2018.02.055

Liu, Xue, de Sherbinin, A., & Zhan, Y. (2019). Mapping Urban Extent at Large Spatial Scales Using Machine Learning Methods with VIIRS Nighttime Light and MODIS Daytime NDVI Data. *Remote Sensing*, *11*(10), 1247. https://doi.org/10.3390/rs11101247

Lloyd, C. T., Chamberlain, H., Kerr, D., Yetman, G., Pistolesi, L., Stevens, F. R., … Tatem, A. J. (2019). Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. *Big Earth Data*, *3*(2), 108–139. https://doi.org/10.1080/20964471.2019.1625151

McGranahan, G., Balk, D., & Anderson, B. (2007). The Rising Tide: Assessing the Risks of Climate Change and Human Settlements in Low Elevation Coastal Zones. *Environment & Urbanisation*, *19*(1), 17–37. https://doi.org/10.1177/0956247807076960

McNeil, D. R., Trussell, T. J., & Turner, J. C. (1977). Spline Interpolation of Demographic Data. *Demography*, *14*(2), 245–252. Retrieved from https://www.jstor.org/stable/2060581

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *Professional Geographer*, *55*(1), 31–42.

Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science2*, *33*, 179–194.

Merriam-Webster. (2019). urban. Retrieved February 7, 2019, from https://www.merriam-webster.com/dictionary/urban

Meyer, P. S., & Ausubel, J. H. (1999). Carrying capacity: A model with logistically varying limits. *Technological Forecasting and Social Change*, *61*(3), 209–214. https://doi.org/10.1016/S0040-1625(99)00022-0

Meyer, W. B., & Turner, B. L. (1992). Human Population Growth and Global Land-Use / Cover Change. *Annual Review of Ecology and Systematics*, *23*(1992), 39–61. https://doi.org/10.2307/2097281

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

Oliveira, S., Oehler, F., San-Miguel-Ayanz, J., Camia, A., & Pereira, J. M. C. (2012). Modeling spatial patterns of fire occurrence in Mediterranean Europe using Multiple Regression and Random Forest. *Forest Ecology and Management*, *275*(July 2012), 117–129. https://doi.org/10.1016/j.foreco.2012.03.003

Openshaw, S. (1984). The modifiable areal unit problem. *Concepts and Techniques* *in Modern Geography*, *38*.

OpenStreetMap Contributers. (2017). OpenStreetMap (OSM) Database. OSM. Retrieved from openstreetmap.org

Parr, J. (2004). The polycentric urban region: A closer inspection. *Regional Studies*, *38*(3), 231–240.

Patel, N., Angiuli, E., Gamba, P., Gaughan, A. E., Lisini, G., Stevens, F. R., … Trianni, G. (2015). Multitemporal Settlement and Population Mapping From Landsat Using Google Earth Engine. *International Journal of Applied Earth Observation and Geoinformation*, *35*(Part B), 199–208. https://doi.org/10.1016/j.jag.2014.09.005

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A. J., Freire, S., Halkia, S., … Syrris, V. (2016). *Operating Procedure for the Production of the Global Human Settlement Layer from Landsat Data of the Epochs 1975, 1990, 2000, and 2014*. Publications Office of the European Union. Retrieved from doi: 10.2788/253582

Pesaresi, M., & Freire, S. (2016). GHS settlement grid, following the REGIO model 2014 in application to GHSL Landsat and CIESIN GPW v4-multitemporal (1975-1990-2000-2015). Retrieved October 26, 2018, from http://data.jrc.ec.europa.eu/dataset/jrc-ghsl-ghs_smod_pop_globe_r2016a

Pesaresi, M., Guo, H., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., … Zanchetta, L. (2013). A Global Human Settlement Layer from Optical HR/VHR Remote Sensing Data: Concept and First Results. *IEEE Journal of Selected Topics in Applied Earth Observation & Remote Sensing*, *6*(5), 2102–2131. https://doi.org/10.1109/JSTARS.2013.2271445

Pontius, R.G., & Millones, M. (2011). Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, *32*(15), 4407–4429. https://doi.org/10.1080/01431161.2011.552923

Pontius, Robert G., Shusas, E., & McEachern, M. (2004). Detecting important categorical land changes while accounting for persistence. *Agriculture, Ecosystems & Environment*, *101*(2–3), 251–268. https://doi.org/10.1016/j.agee.2003.09.008

Potere, D, & Schneider, A. (2007). A critical look at representations of urban areas in global maps. *GeoJournal*, *69*(1–2), 55–80. https://doi.org/10.1007/s10708-007-9102-z

Potere, David, Schneider, A., Angel, S., & Civco, D. (2009). Mapping urban areas on a global scale: which of the eight maps now available is more accurate? *International Journal of Remote Sensing*, *30*(24), 6531–6558. https://doi.org/10.1080/01431160903121134

Pozzi, F., & Small, C. (2005). Analysis of urban land cover and population density in the United States. *Photogrammetric Engineering & Remote Sensing2*, *71*, 719–726.

R Core Team. (2016). R: A Language and Environment Layer for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org

Rogan, W. J., & Gladen, B. (1978). Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, *107*(1), 71–76.

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *PLoS One*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

Sante, I., Garcia, A. M., Miranda, D., & Crecente, R. (2010). Cellular Automata Models for the Simulation of Real-world Urban Processes: A Review and Analysis. *Landscape and Urban Planning*, *96*, 108–122. https://doi.org/10.1016/j.landurbplan.2010.03.001

Schneider, A., Friedl, M. A., McIver, D. K., & Woodcock, C. E. (2003). Mapping Urban Areas by Fusing Multiple Sources of Coarse Resolution Remotely Sensed Data. *Photogrammetry & Remote Sensing*, *69*(12), 1377–1386.

Schneider, A., Friedl, M. A., & Potere, D. (2010). Mapping Urban Areas Using MODIS 500-m Data: New Methods and Datasets Based on "Urban Ecoregions." *Remote Sensing of the Environment*, *114*, 1733–1746.

Schneider, A., & Woodcock, C. E. (2008). Compact, Dispersed, Fragmented, Extensive? A Comparison of Urban Growth in Twenty-five Global Cities using Remotely Sensed Data, Pattern Metrics and Census Information. *Urban Studies*, *45*(3), 659–692. https://doi.org/10.1177/0042098007087340

Seto, K. C., Fragkias, M., Guneralp, B., & Reilly, M. K. (2011). A Meta-Analysis of Global Urban Land Expansion. *PLoS One*, *6*(8), e23777. https://doi.org/10.1371/journal.pone.0023777

Seto, K. C., Guneralp, B., & Hutyra, L. R. (2012). Global Forecasts of Urban

Expansion to 2030 and Direct Impacts on Biodiversity and Carbon Pools. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(40), 16083–16088. https://doi.org/10.1073/pnas.1211658109

Shi, K., Huang, C., Yu, B., Yin, B., Huang, Y., & Wu, J. (2014). Evaluation of NPP-VIIRS night-time light composite data for extracting built-up urban areas. *Remote Sensing Letters*, *5*(4), 358–366. https://doi.org/10.1080/2150704X.2014.905728

Small, C. (2009). The color of cities:An overview of urban spectral diversity. In M. Herold & P. Gamba (Eds.), *Global Mapping of Human Settlements* (pp. 59–106). New York: Taylor & Francis.

Small, C., Elvidge, C. D., Balk, D., & Montgomery, M. (2011). Spatial scaling of stable night lights. *Remote Sensing of Environment*, *115*, 269–280. https://doi.org/10.1016/j.rse.2010.08.021

Small, C., Pozzi, F., & Elvidge, C. D. (2005). Spatial analysis of global urban extent from DMSP-OLS night lights. *Remote Sensing of Environment*, *96*, 277–291. https://doi.org/10.1016/j.rse.2005.02.002

Sorichetta, A., Hornby, G. M., Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). High-resolution gridded population distribution datasets of Latin America in 2010, 2015, and 2020. *Scientific Data*, *2*, 150045. https://doi.org/10.1038/sdata.2015.45

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

Stephenson, J., Newman, K., & Mayhew, S. (2010). Population dynamics and climate change: What are the links? *Journal of Public Health*, *32*(2), 150–156. https://doi.org/10.1093/pubmed/fdq038

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-sensed Data and Ancillary Data. *PLoS One*, *10*(2), e0107042. https://doi.org/10.1371/journal.pone.0107042

Sverdlik, A. (2011). Ill-health and poverty: A literature review on health in informal settlements. *Environment and Urbanisation*, *23*(1), 123–155. https://doi.org/10.1177/0956247811398604

Tayyebi, A., Pekin, B. K., Pijanowski, B. C., Plourde, J. D., Doucette, J. S., & Braun, D. (2013). Hierarchical modeling of urban growth across the conterminous

USA: Developing meso-scale quantity drivers for the Land Transformation Model. *Journal of Land Use Science*, *8*(4), 422–442. https://doi.org/10.1080/1747423X.2012.675364

Tobler, W., Deichmann, U., Gottsegen, J., & Maloy, K. (1997). World Population in a Grid of Spherical Quadrilaterals. *International Journal of Population Geography*, *3*, 203–225.

U.N. Enviroment Programme World Conservation Monitoring Centre, & IUCN World Commission on Protected Areas. (2015, November 7). World Database on Protected Areas. IUCN & UNEP. Retrieved from https://www.protectedplanet.net/

UCL Geomatics. (2017). *Land Cover CCI Product User Guide Version 2.0*. Retrieved from http://maps.elie.ucl.ac.be/CCI/viewer/download/ESACCI-LC-Ph2-PUGv2_2.0.pdf

United Nations. (2015a). *World Population Prospects: The 2014 Revision*. Washington, D. C.

United Nations. (2015b). *World Urbanisation Prospects: The 2014 Revision* (World Urbanisation Prospects No. ST/ESA/SER.A/366). New York, New York: United Nations, Dept. of Economic and Social Affairs, Population Division. Retrieved from https://population.un.org/wup/Publications/Files/WUP2014-Methodology.pdf

United Nations. (2016). *Transforming Our World: The 2030 Agenda for Sustainable Development*. Retrieved from https://sustainabledevelopment.un.org/content/documents/21252030 Agenda for Sustainable Development web.pdf

United Nations. (2018). *World Urbanisation Prospects: The 2018 Revision*. New York. Retrieved from https://population.un.org/wup/Publications/Files/WUP2018-Methodology.pdf

Utazi, C. E., Thorley, J., Alegana, V. A., Ferrari, M. J., Takahashi, S., Metcalf, C. J. E., … Tatem, A. J. (2018). High resolution age-structured mapping of childhood vaccination coverage in low and middle income countries. *Vaccine*, *36*(12), 1583–1591. https://doi.org/10.1016/j.vaccine.2018.02.020

Venables, W. N., & Ripley, B. D. (2002). *Modern Applied Statistics with S* (4th ed.). New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Verburg, P. H., de Koning, G. H. J., Kok, K., Veldkamp, A., & Bouma, J. (1999). A spatial explicit allocation procedure for modelling the pattern of land use change based upon actual land use. *Ecological Modelling*, *116*, 45–61.

Verburg, P. H., Schot, P. P., Dijst, M. J., & Veldkamp, A. (2004). Landuse Change Modelling: Current Practice and Research Priorities. *GeoJournal*, *61*, 309–324.

Verburg, P. H., Soepboer, W., Veldkamp, A., Limpiada, R., Espladon, V., & Mastura, S. S. A. (2002). Modeling the Spatial Dynamics of Regional Land Use: The CLUE-S Model. *Environmental Management*, *30*(3), 391–405. https://doi.org/10.1007/s00267-002-2630-x

Von Thunen, J. H. (1966). *Von Thunen's "Isolated State": An English Translation of "Der Isolierte Staat."* (C. M. Wartenberg & P. Hall, Eds.). Oxford, UK: Pergamon Press.

Weiss, D. J., Nelson, A., Gibson, H. S., Temperley, W., Peedell, S., Lieber, A., … Gething, P. W. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, *553*(7688), 333–336. https://doi.org/10.1038/nature25181

White, R., & Engelen, G. (1997). Cellular Automata as the Basis of Integrated Dynamic Regional Modelling. *Environment and Planning B*, *24*, 235–246.

White, R., & Engelen, G. (2000). High Resolution Modelling of the Spatial Dynamics of Urban and Regional Systems. *Computers, Environment, and Urban Systems*, *24*(383–400).

Wicht, M., & Kuffer, M. (2019). The continuous built-up area extracted from ISS night-time lights to compare the amount of urban green areas across European cities. *European Journal of Remote Sensing*, *52*(sup2), 58–73. https://doi.org/10.1080/22797254.2019.1617642

WorldPop, S. of G. and E. S. U. of, Department of Geography and Geosciences, U. of L., Département de Géographie, U. de N., & Center for International Earth Science Information Network (CIESIN), C. U. (2018). Global High Resolution Population Denominators Project. Bill and Melinda Gates Foundation (OPP1134076). https://doi.org/10.5258/SOTON/WP00644

Zelinsky, W. (1971). The Hypothesis of the Mobility Transition. *Geographical Review*, *61*(2), 219–249.

Zhang, Q., Pandey, B., & Seto, K. C. (2016). A Robust Method to Generate a Consistent Time Series From DMSP/OLS Nighttime Light Data. *IEEE Transactions on Geoscience and Remote Sensing*, *54*(10), 5821–5831.

https://doi.org/10.1109/TGRS.2016.2572724

Zhang, Q., & Seto, K. C. (2011). Mapping urbanisation dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data. *Remote Sensing of Environment*, *115*(9), 2320–2329. https://doi.org/10.1016/j.rse.2011.04.032

Chapter 4 Predicting near-future built-settlement expansion using relative changes in small area populations

Nieves, J. J., Bondarenko, M., Sorichetta, A., Steele, J. E., Kerr, D., Carioli, A., Stevens, F. R., Gaughan, A. E., & A. J. Tatem. "Predicting near-future built-settlement expansion using relative changes in small area populations"

The version included in this thesis is the author's version of a work that was submitted for publication in *Remote Sensing*. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication. A definitive version was subsequently published in *Remote Sensing* 12(10):1545

https://doi.org/10.3390/rs12101545

remote sensing

Article

Predicting Near-Future Built-Settlement Expansion Using Relative Changes in Small Area Populations

Jeremiah J. Nieves [1,*], Maksym Bondarenko [1], Alessandro Sorichetta [1], Jessica E. Steele [1], David Kerr [1], Alessandra Carioli [1], Forrest R. Stevens [1,2], Andrea E. Gaughan [1,2] and Andrew J. Tatem [1]

[1] WorldPop, School of Geography and Environmental Science, University of Southampton, Southampton SO17 1BJ, UK; M.Bondarenko@soton.ac.uk (M.B.); A.Sorichetta@soton.ac.uk (A.S.); js1m14@soton.ac.uk (J.E.S.); dkerr83@gmail.com (D.K.); alessandracarioli@gmail.com (A.C.); forrest@forreststevens.com (F.R.S.); ae.gaughan@louisville.edu (A.E.G.); A.J.Tatem@soton.ac.uk (A.J.T.)

[2] Department of Geography and Geosciences, University of Louisville, Louisville, KY 40222, USA *
Correspondence: jeremiah.nieves@outlook.com

Received: 17 March 2020; Accepted: 11 May 2020; Published: 12 May 2020

Abstract: Advances in the availability of multi-temporal, remote sensing-derived global built-/human-settlements datasets can now provide globally consistent definitions of "human-settlement" at unprecedented spatial fineness. Yet, these data only provide a time-series of past extents and urban growth/expansion models have not had parallel advances at high-spatial resolution. Here our goal was to present a globally applicable predictive modelling framework, as informed by a short, preceding time-series of built-settlement extents, capable of producing annual, near-future built-settlement extents. To do so, we integrated a random forest, dasymetric redistribution, and autoregressive temporal models with open and globally available subnational data, estimates of built-settlement population, and environmental covariates. Using this approach, we trained the model on a 11 year time-series (2000–2010) of European Space Agency (ESA) Climate Change Initiative (CCI) Land Cover "Urban Areas" class and predicted annual, 100m resolution, binary settlement extents five years beyond the last observations (2011–2015) within varying environmental, urban morphological, and data quality contexts. We found that our model framework performed consistently across all sampled countries and, when compared to time-specific imagery, demonstrated the capacity to capture human-settlement missed by the input time-series and the withheld validation settlement extents. When comparing manually delineated building footprints of small settlements to the modelled extents, we saw that the modelling framework had a 12 percent increase in accuracy compared to withheld validation settlement extents. However, how this framework performs when using different input definitions of "urban" or settlement remains unknown. While this model framework is predictive and not explanatory in nature, it shows that globally available "off-the-shelf" datasets and relative changes in subnational population can be sufficient for accurate prediction of future settlement expansion. Further, this framework shows promise for predicting near-future settlement extents and provides a foundation for forecasts further into the future.

Keywords: Urban; growth model; forecast; built; settlement; machine learning; time series

1. Introduction

In 2018, 55 percent of the world's population lived in urbanized areas, but this is projected to increase to 68 percent by 2050, due to natural population growth, continued rural to urban migration, and the conversion of rural to urban land [1–3]. Most of this anticipated urban growth will be in low and middle-income countries, specifically in small to medium sized settlements, where the majority of urban populations reside [1,4]. Logically, this growth, in conjunction with climate change, presents questions regarding sustainable development. Answers to these questions are dependent upon better understanding past and current urbanization trends to better predict future trends, minimize potential adverse outcomes and environmental impact, and maximize the benefits that can come from urbanization [1,4–6]. Accordingly, there is a continued need for globally comparable and standardized urban environment datasets and projections [4,6,7]. Particularly as internationally coordinated and global efforts for sustainable development, such as under the Sustainable Development Goals [8], are undertaken. The provision of these data needs to be transparent, sustainable, comparable across space and time, and available to all while being able to cope with the many definitions of urban, e.g. administrative-based, Remote Sensing (RS)-based, or population-based definitions [8–10].

While detailed and regular data on urban areas often exists within high-income countries, middle and lower-income countries often lack these data, use country specific definitions of urban, or have data that is not easily accessible. Often, practitioners turn to RS-based global data that has consistently extracted urban areas and features using a definition based upon the observable human, built land cover. Recent advances have produced globally consistent urban feature datasets, which maintain relatively high spatial resolution/fidelity (<= 100m) while capturing smaller/less-dense/more-fragmented settlements [11–15]. Specifically, the availability of urban feature datasets globally capturing areas of Built-Settlement (BS), above ground structures that can support human habitation and or related economic processes [12,16,17], have become more common, e.g. [12–15,18–20]. However, these datasets still have limited temporal resolution, i.e. single time observation or cross-sectional with many years between observations. Increased temporal coverage is desirable but sacrificing spatial resolution to do so is problematic as most human settlements, particularly those in low- to middle-income countries, are relatively smaller and less densely developed [1,21–23]. Compounding this is the typical time lag between global image acquisitions and the resulting dataset of built-settlement or, more generally, urban features and the associated processing costs. Further, some datasets are only produced once or cease updating with additional observations in time, leaving users of the data without continued support for a dataset-specific definition of urban. Hereafter, we refer to the general concept of "urban" as such, and use the "built-environment" to refer to all areas characterized by the presence of anthropogenic features, and use "urban features" to refer to *objects* within the built-environment, e.g. roads, buildings, parks. Specifically, the scenario of needing to project built-settlement extent data past last observations would logically propose extrapolative modelling as a solution.

To this regard, it is worth highlighting that the majority of the literature and existing models for projections of urban and built-environment growth focus on North America, Europe, and China, with many being city/area/regionally specific [24]. Furthermore, many of the existing continental- and global-extent urban future growth models are solely meant for exploring potential future scenarios as opposed to projecting near future urban growth grounded upon local contemporary and past observed dynamics [25–27]. Other models are produced from city- or country-level samples, datasets with substantial definitional or spatial/temporal disagreement, or utilise arbitrary thresholds without validation for determining non-urban-to-urban area conversion [3,28–31] Of these, many are not driven by subnational variations to determine larger scale dynamics of urban growth and transition distributions, e.g. they are statistically "global" models. Some models do not output explicit spatial extents, e.g. country-level totals of projected urban area, limiting their utility [3]. Together, these issues combined indicate a need for methods to produce a flexible and robust method of generating spatially explicit regular time series of predicted future urban environment expansion across the globe.

Our goal was to leverage developments in statistical methods, data availability, and computing resources to create a globally applicable urban expansion modelling framework to project beyond the last observations while addressing the above existing needs. Using observed time-series of BS extents and coincident small area population changes we were able to produce spatially explicit annual BS extent maps representing projected near-future BS expansion for multiple points in time. Here we introduce such a modelling framework and validate its performance against withheld time-specific past RS-derived observations and time-specific manual delineations of BS.

2. Materials and Methods

2.1. Study Areas and Data

To test across a variety of BS morphologies, environmental contexts, and developmental contexts, in addition to countries with varying spatial details of the input census-based population data, we sample countries less present in previous spatial urban and BS modelling studies [24], including Switzerland, Panama, Uganda, and Vietnam (Table 1). Additionally, these countries were chosen to capture a variety of population magnitudes, densities, and distributions across space as well as socio-economic, urban morphological, topographical, and data quality (e.g. spatial fineness of subnational population data) contexts. Given that this extrapolative framework builds off the previously fit interpolative Built-Settlement Growth Model (BSGMi) [16], the same set of covariates were used as in [16] for either predicting transition probability in the random forest (Table 2, superscript “c”) or in the remainder of the disaggregative process. These covariates were selected based upon previous literature to give immediate environmental context and information regarding settlement connectivity and proximity [28,32,33], e.g. negative relationship between slope and likelihood of transition, positive relationship between likelihood of transition and distance to existing BS. Covariates were time specific or assumed to be temporally invariant (Table 2), and were pre-processed and appropriately resampled to 3 arc seconds (~ 100m at the Equator) as detailed in Lloyd et al. [34].

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. Adapted from Nieves et al. [16].

Country	Average Spatial Resolution [a]	Period	Initial Non-Built Area (pixels)	Period Transition Prevalence [b]
Panama	10.9 km	2000–2010	8,901,004	0.12%
		2010–2015	8,890,339	0.75 %
Switzerland	3.9 km	2000–2010	6,816,510	1.64%
		2010–2015	6,704,973	0.01%
Uganda	12.2 km	2000–2010	28,231,555	0.11%
		2010–2015	28,200,084	0.04 %
Vietnam	21.7 km	2000–2010	40,108,425	0.11% 0.29%
		2010–2015	39,990,858	

[a] Average spatial resolution is the square root of the average subnational area, in km, and can be thought of as analogous to pixel resolution with smaller values indicating finer areal data and vice versa [35]

[b] Note: the Switzerland data suffered from disproportionate, relative to manually interpreted 30 cm true-colour imagery, amounts of growth as indicated by the European Space Agency (ESA) Remote Sensing (RS)-derived extents between 2000–2005 and is thought by Nieves et al. [16] to be due to the 2003–2004 shift from delineating land cover changes at 300 m to using imagery to delineate at 150 m, in conjunction with the highly variable terrain in Switzerland compounding classification attempts.

Table 2. Data used for estimating the annual number of non- Built-Settlement (BS) to BS transitions at the unit level (i.e. demand quantification), predicting the pixel level probability surface of those transitions, and performing the spatial allocation procedures of the model. Adapted from Nieves et al. [16].

Covariate	Description	Use b, d	Time Point(s)	Original Spatial Resolution	DataSource(s)
Built-settlement [b]	Binary BS extents	Demand Quantification Spatial Allocation	2000–2010	10 arc sec	[36]
Distance To nearest Edge (DTE) of Built-settlement	Distance to the nearest BS edge	Spatial Allocation [c]	2000, 2010	10 arc sec	[36]
Proportion Built-settlement 1,5,10,15	Proportion of pixels that are BS within 1,5,10, or 15-pixel radius	Spatial Allocation [c]	2000,2010	10 arc sec	[36]
Elevation	Elevation of terrain	Spatial Allocation [c]	2000; Time Invariant	3 arc sec	[37]
Slope	Slope of terrain	Spatial Allocation [c]	2000; Time Invariant	3 arc sec	[37]
DTE Protected Areas Category 1	Distance to the nearest level 1 protected area edge	Spatial Allocation [c]	2010	Vector	[34,38]
Water	Areas of water	Restrictive Mask		5 arc sec	[34,39]
Subnational Population	Annual population by sub-national units	Demand Quantification	2000–2020	Vector	[40]
Weighted Lights-at-Night (LAN) [d]	Annual lagged and sub-national unit normalised LAN	Spatial Allocation [d]	2000–2016	30 arc sec (2000-011) 15 arc sec (2012-016)	DMSP [34,41] VIIRS [34,42]
Travel Time 50k	Travel time to the nearest city centre containing at least 50,000 people	Spatial Allocation [c]	2000	30 arc sec	[34,43]
ESA CCI Land Cover (LC) Class [a]	Distance to nearest edge of individual land cover classes	Spatial Allocation [c]	2000, 2010	10 arc sec	[34,36]
Distance to OpenStreetMap (OSM) Rivers	Distance to nearest OSM river feature	Spatial Allocation [c]	2017	Vector	[34,44]
Distance to OpenStreetMap (OSM) Roads	Distance to nearest OSM road feature	Spatial Allocation [c]	2017	Vector	[34,44]
Average Precipitation	Mean Precipitation	Spatial Allocation [c]	1950–2000	30 arc sec	[34,45]
Average Temperature	Mean temperature	Spatial Allocation [c]	1950–2000	30 arc sec	[34,45]

a Some land cover classes were collapsed prior to calculating distance to edge: 10–30 → 11; 40–120 → 40; 150–153 → 150; 160–180 → 160 (Sorichetta et al>, 2015)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 (see c) 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 Used as predictive covariates in the random forest calculated probabilities of transition d See Nieves et al. [16] for details on the construction of weighted LAN

2.1.1. Built-Settlement Data

Our chosen representation of BS was the "Urban" class, number 190, of the annual European Space Agency Climate Change Initiative thematic land cover dataset (https://www.esa-landcover-cci.org/; hereafter, ESA). We selected the ESA RS-derived extents data for its annual coverage, at the time of the study, from 1992 to 2015. It has recently been extended to provide coverage for the years 2016–2018 [46]. While ESA RS-derived extents have moderate spatial resolution, 10 arc sec resolution (~ 300m at Equator), its annual temporal resolution allows for the withholding of years for validation. In our period of interest, 2000 to 2015, the ESA data begins with a Medium Resolution Imaging Spectrometer (MERIS) imagery derived baseline land cover map and detects thematic class changes from this map using 30 arc second (~ 1 km at the Equator) SPOT VGT imagery (1999–2013) and PROBA-V imagery (2014–2015) [47]. Any detected changes observed over two or more years are delineated at 30 arc second resolution, if prior to 2004, and, beginning with 2004, are further delineated at 10 arc second resolution using the higher resolution MERIS or PROBA-V imagery [47]. Specific to the "Urban" class, ESA incorporates the Global Human Settlement Layer (GHSL) [12,18] and Global Urban Footprint (GUF) [13] datasets to better define the class and integrate elements of two BS datasets within the overall thematic built-environment context. Initial validation efforts estimate the 2015 "Urban" class user and producer accuracies between 86–88 percent and 51–60 percent, respectively, but no information on the other years currently exist [47].

While ESA utilises the term "urban", it is more correctly capturing aspects of the built environment. Given the integration of the GHSL and GUF data sets, which capture built-settlement, into the ESA "urban" class, we have reason to believe that the ESA "urban" class is more correctly operating on a functional definition of "built-settlement" or "built-settlement"-like, and refer to it as such. For a more detailed discussion on built-settlement and remote-sensing representations, readers are referred to Nieves et al. [16].

2.1.2. Population Data

Annual subnational unit area (hereafter simply "unit,") population estimates, for 2000 through 2020, were based upon the Gridded Population of the World version 4 (GPWv4) input data [40] were produced by the Center for International Earth Science Information Network (CIESIN) and spatially harmonized as described in Lloyd et al. [34]. To clarify, we are not using the gridded GPW product, which has uniform population density within a given unit, but we are using the same tabular population count data and the associated unit areas. These counts are based upon censuses/official estimates, interpolated at the subnational level per [40] to obtain annual estimates. Each unit possesses a unique ID, referencing a globally consistent grid (3 arc seconds), with the unit areas having globally harmonized coastlines and international borders. It is worth noting that the population count data utilised here are not adjusted to the U.N. country total population estimates, which are used to account for potential biases and errors. Further, the two primary sources of uncertainty in this dataset are linked to the census figures/official estimates and the simple regression used to obtain the annual estimates with few assumptions.

2.1.3. OpenStreetMap Data

OpenStreetMap (OSM) is an open database of user-contributed, edited, and curated spatial data also known as. While OSM offers global extent, like other Volunteered Geographic Information (VGI) [48], its completeness varies across space, with particular gaps in low and middle income countries, and has data quality that can vary both within and between countries [49,50]. Contrastingly, in the best of cases, OSM can approach the quality of official datasets [51]. However, agreed upon means of assessing VGI data quality and accuracy varies and is still debated [52]. Nonetheless, OSM data are used to fill data gaps where official/commercial datasets do not exist or are not publicly accessible and have improved or produced useful analyses and derived datasets, (e.g. [13,34,53–57]).

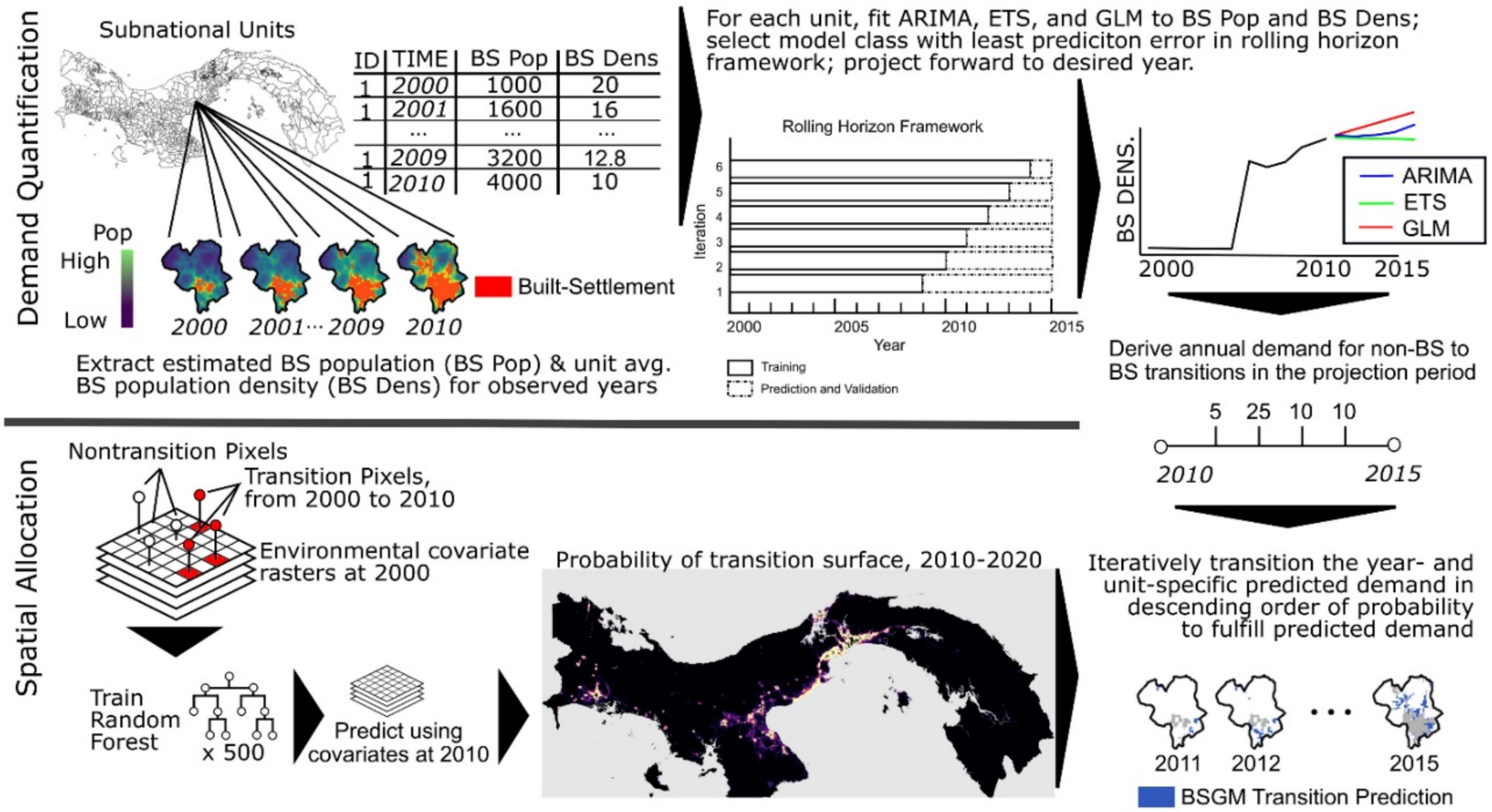
For validation, we utilised the OSM building footprints around the municipalities of Visp, Brig-Glis, Naters, and Ried-Brig, Switzerland, where agreement between modelled extents and RS-derived extents were particularly large. The mountainous 119 km^2 area (rectangular bounds: 7.8606508° 46.2779033°; 8.0224478°, 46.3298123°) had a 2015 combined mid-year population of approximately 32,430 [58]. It contained 8,083 buildings manually delineated by OSM contributors, of which we contributed over an additional 1,700 buildings in an effort to have near 100 percent coverage of permanent vertical structures covered by the definition of BS. We inspected all building footprints in the area for accuracy and temporal coincidence with true colour imagery in 2015. The resource intensive nature of manually delineating and checking building footprints precluded us from carrying out more widespread validations of this nature during this study. The building footprints are provided in the linked data repository (https://data.mendeley.com/datasets/cm6bnzvzfj/1).

2.2. Built-Settlement Growth Model extrapolation (BSGMe)

2.2.1. Overview

Here we take annual time-series of BS extents spanning 2000–2010 and estimated annual changes in BS population and unit-average BS population density changes to predict short-term (within five years) BS extents from 2011 through 2015. BS population is the population coincident with the BS extents and unit-average BS population density is the BS population of a unit divided by the BS area within the same unit. We refer to the set of years making each time series as *TS* where *TS* = {*2000, 2001, … ,2010*} and, expanding the notation from Nieves et al. [16], the first and last years of the input time series are referred to as *t0* and *t1*, respectively. We test this extrapolative Built-Settlement Growth Model (BSGMe) framework using an annual time series of RS-based ESA BS extents from 2000–2010 (TS_{ESA}).

Similar to the BSGMi framework [16], the BSGMe framework has two primary components of "Demand Quantification" and "Spatial Allocation", shown here in Figure 1.



ID	TIME	BS Pop	BS Dens
1	2000	1000	20
1	2001	1600	16
	...	...	...
1	2009	3200	12.8
1	2010	4000	10

Extract estimated BS population (BS Pop) & unit avg. BS population density (BS Dens) for observed years

For each unit, fit ARIMA, ETS, and GLM to BS Pop and BS Dens; select model class with least prediciton error in rolling horizon framework; project forward to desired year.

Derive annual demand for non-BS to BS transitions in the projection period

5
25
10
10
2010
2015
Spatial Allocation
Nontransition Pixels
Transition Pixels, from 2000 to 2010
Environmental covariate rasters at 2000
Train Random Forest
x 500
Predict using covariates at 2010
Probability of transition surface, 2010-2020

Iteratively transition the year- and unit-specific predicted demand in descending order of probability to fulfill predicted demand

Figure 1. High-level generalization of the Built-Settlement Growth Model extrapolation (BSGMe) modelling framework when predicting for short-term Built-Settlement (BS) expansion. Note, example maps and numbers are not to scale. Figure modified from [16].

We generalize the BSGMe framework with following steps:

1. Create gridded population maps for each year in the input *TS*, following Stevens et al. [54].
2. For all years in the *TS*, extract the unit-specific population sum that is coincident with the year's corresponding BS extents and derive the unit-average BS population density
3. Independently for each unit, and using a rolling origin validation, select the single best fitting model for BS population and, separately, unit-average BS population density from three classes of models:

    - Auto-Regressive Integrated Moving Average (ARIMA),
    - Error, Trend, Seasonality (ETS), and
    - Generalized Linear Model (GLM) given log-transformed inputs.

4. For each unit, use the final selected model for BS population and for unit-average BS population density to predict short-term annual BS population and annual unit-average BS population density starting with year *t1+1* and ending with year *t1+h*, where in this case 1 ≤ *h* ≤ 5 and represents the projection horizon, in numbers of years.
5. Use these estimates to derive the unit-specific annual quantity demand of non-BS-to-BS transitions by dividing the BS population by the BS population density.
6. Create a transition probability surface using a Random Forest (RF) based upon the observed transitions between *t0* and *t1* of the input time-series and covariates corresponding to *t0*.
7. Take the fit relationships between the occurrence of transitions and the predictive covariates, contained in the final RF model, and predict the future non-BS-to-BS transition probability surface using the same covariates, but corresponding to year *t1*, as the input.
8. For each unit and iteratively for all years *t1+1* through *t1+h*, spatially disaggregate the predicted annual unit-level transitions (steps 1–5) using the base transition probability surface (steps 5–6) and, if available, unit-relative weights derived from changes in lights-at-night brightness, similar to Nieves et al. [16].

These steps produce annual binary spatial predictions of BS extent in gridded format. All modelling and analyses were carried out using R 3.4.2 [59] and utilised the IRIDIS 4 high-performance computing cluster. All code is provided in the linked data repository (https://data.mendeley.com/datasets/cm6bnzvzfj/1). Full process diagrams are provided in Appendix A, Figures A1 and A2.

2.2.2. Demand Quantification

Built-Settlement Population Estimation

To obtain a set of annual estimated population surfaces for our study areas, we used the method detailed by Stevens et al. [54] to dasymetrically disaggregate [60,61] the census-based population from the unit-level to 3 arc second (~100m at the Equator) pixels. We independently modelled each country and year utilising time-specific and, assumed, time-invariant predictive covariates (see Appendix A, Table A1). We included the distance-to-nearest BS edge at the year 2000 and the distance-to-nearest

BS edge for the given year as predictive covariates. This corresponded with our assumption that population relates to inner parts of BS agglomerations differently from the outer parts and to avoid exaggerated areas of low population density relative to previously modelled years [16,62]. Annually, for each unit, we extracted and summed the populations from pixels that were within year-specific BS extents and derived the annual unit-average BS population density. This resulted in annual time-series of BS population estimates and unit-average BS population densities for every unit in the study area, covering eleven years.

Time-Series Model Fitting and Built-Settlement Population Projections

Using these annual unit-level time-series, we predicted future unit BS population and unit-average BS population density using a single model fitting and selection process detailed in Figure 1. For each unit, this process fits three classes of models: ARIMA models, ETS models, and an identity-link GLM model with log-transformed input values, all using a rolling origin framework validation in the final, i.e. between-class, model selection process.

ARIMA models [63–65] and ETS models [64–67] are two autoregressive model classes often applied to time-series data, including population forecasts [68]. Both classes have dependent model terms based upon preceding values in the input time-series. ETS models are based upon the assumption of non-stationary, i.e. the mean and variance of the underlying process are not constant, and can approximate non-linear processes [64]. Conversely, ARIMA models assume stationarity and a linear correlation between the values of the time-series, but remain a standard in forecasting time-series [64,69]. The best model within the ARIMA class relies on an automated fitting procedure utilising unit root tests, iterative step-wise parameter fitting, and the resultant lowest Akaike Information Criterion (AIC) value, as described in detail by Hydman and Khandakar [64]. ETS class models are selected in an automated fashion, as described in Hyndman et al. [65], utilising maximum likelihood parameter estimation, the corrected AIC (AICc), and bootstrapping simulation. For the ARIMA and ETS model classes only the number of years since year *t0* and temporally preceding values in the input time-series were available as predictive covariates.

Generalized Linear Models (GLMs) provide a single consistent framework for linking the linear-based systematic elements of regression-type models, associated with Normal, binomial, Poisson, gamma, and other statistical distributions, with their respective random components through an integrated fitting procedure based upon maximum likelihood [70]. Here, we utilised an identity link function, and provided log-transformed input data with the number of years since year *t0* as the sole predictive covariate.

During the fitting of these model classes we utilised a rolling origin validation (Figure 2) of each model class in anticipation of needing to determine the final model based upon a single metric of error across the different number of years predicted into the future. A rolling origin validation fits a selected model upon an iteratively changing sample size and an inversely changing number of future time steps, i.e. "the rolling origin" (Figure 2) [71–73]. We used the Median Absolute Percent Error (MDAPE) as our forecasting error metric as opposed to the more common Mean Absolute Percent Error (MAPE). The MAPE, compared to other metrics, has the advantage of avoiding large errors when the true value is near zero [74]. The MDAPE retains the advantages of the MAPE but is less influenced by extreme values and is more robust than the MAPE [69,74]. It can be written as:

$$MDAPE = median\left(\left|\frac{\hat{y}-y}{y} * 100\right|\right) \quad (1)$$

where $y\hat{}$ is the predicted outcome of interest and y is the withheld observed outcome.

Given our short input time-series (N_{ts} = 11) and our projection horizon between one and five years ($1 \leq h \leq 5$), we utilised a maximum horizon of five years in the model fitting too. This meant the model classes were iteratively fit with between six (i.e. 2000–2005) and ten (i.e. 2000–2009) input observations, with all other observations withheld, and then predicted between one and five years, respectively, forward of the last input year of the given iteration sample. Each iteration produced a set of annual absolute percent errors for the projected years, of which the median was recorded. The sum of MDAPE values across all iterations represents the total error of each model class for the given unit. Written mathematically, for a given unit *i*, maximum horizon length *h*, and *a* being the index of the given set of iterations, the MDAPE sum within the rolling origin framework can be written as

$$MDAPE_{i_{sum}} = \sum_{a=0}^{h}[MDAPE_a] = \sum_{a=0}^{h}\left[median(\left\{\left|\frac{\hat{y}_k-y_k}{y_k} * 100\right|\right\}_{k=n_{ts+1}}^{n_{ts+h}})\right] \quad (2)$$

where the sample training series for a given iteration can be written as $n_{ts} = t1 + a - h$ and the set of projected years within an iteration are calculated for each year *k* that takes on values between $n_{ts+1}, \dots, n_{ts+h}$, e.g. for $h = 3$ and $a = 3$ the models are fit on years 1 to 8 with a set of predictions made for $\{\hat{y}_{n_{ts}+1}, \hat{y}_{n_{ts}+2}, \hat{y}_{n_{ts}+3}\}$. After the rolling origin framework finished, for each unit, we selected the best model between model classes based upon the lowest $MDAPE_{isum}$ and fit the selected model class on the entire available time series. Normally, using the entire time series is cause for concern of model over fitting. However, our larger concern was that excluding later observations in the extremely short time series could lead to excluding important information late in the series. Therefore, we assumed that fitting only on a subset of the time-series would be as harmful, or more so, than potentially overfitting any given unit. After the refitting, and independently for each unit, we predicted the final outcome of interest through our projection horizon, in this case 2011–2015. Full process diagram of this sub-procedure is provided in Appendix A, Figure A2.

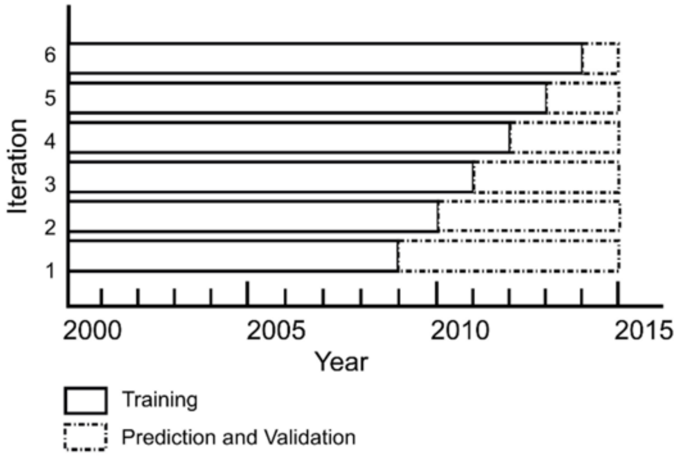


Figure 2. Unit-level model fitting process for fitting and selecting the final model, between three classes of models, used to predict short-term future BS population and future unit-average BS population density. Here we employ a rolling origin framework, with the final

model selected based upon the smallest sum of the Median Absolute Percent Error (MDAPE).

This time-series model selection and prediction procedure was used twice in the demand quantification component of the BSGMe: once for predicting future BS population and once for predicting future unit-average BS population density (Figure 1). For predicting future BS population, we first transformed, and later back-transformed, BS population to an "BS/Non-BS Ratio" to ensure BS population never exceeded total population [1]. We then calculated the final year and unit-specific number of projected non-BS-to-BS transitions by dividing the projected BS population by the corresponding projected BS population density.

2.2.3. Spatial Allocation

Projecting non-Built-Settlement (BS)-to-BS Transition Probabilities Surface

After calculating annual unit-level demand for non-BS-to-BS transitions, we spatially allocated transitions to the pixel level, producing annual projected BS extents. First, we trained a RF on transitions observed between 2000 and 2010 with spatially coincident covariates corresponding to the year 2000 (Table 2). This RF was created following the sampling and training procedures in Nieves et al. [16] where an iterative covariate selection procedure was employed, removing covariates that did not improve the accuracy of the RF model. In this scenario, we were assuming that relationships observed between transitions and the predictive covariates persist into the near future. Therefore, we projected forward to estimate the probability of transition surface after 2010 by using 2010 representative covariates as input covariates (Figure 1). The values of the resulting probability surface range from 0.00 to 1.00 and represent the posterior probability of a pixel being classified as transitioning between, originally 2000 and 2010, 2010 and 2015 [75]. We elected to use a RF due to its efficiency and scalability as well as its ability to model complex interactions and non-linear phenomenon using a non-parametric approach with minimal input [75]. Further, RFs have been shown in at least one study to outperform other machine learning type methods, such as support vector machines [76], and showed satisfactory performance in Nieves et al. [16].

Annually Adjusting non-BS-to-BS Transition Probabilities

While many projections are "truly future" scenarios and no earth observation data would be available, here we are validating the framework within a scenario where the "future" projection period is one where the input BS extent dataset does not have coverage, i.e. as if ESA had stopped producing the dataset at 2010, and we have access to observed lights-at-night (LAN) data during our projection period (2011–2015). With this, we follow the procedure in Nieves et al. [16] of using average annual unit-normalized lagged LAN brightness to modify the period probability produced by the RF to a more annual representation of the unit-specific probabilities of transition. The assumption behind this process is that pixels with larger unit-relative changes in annual LAN brightness correspond to a larger probability of non-BS-to-BS transition occurring at those location and vice versa. That is, if a relatively large increase, with respect to the given subnational unit, in the LAN brightness occurred between years and given that the area was not already BS, we would assume this corresponded to a higher probability of non-BS-to-BS transition having occurred.

Using these annually adjusted unit-relative probabilities, we followed the procedure in Nieves et al. [16] to spatially disaggregate the demand quantification component-derived projected annual transitions from the unit-level to the pixel-level (Figure 1). Differing from Nieves et al. [16], we did not restrict where the transitions can occur, excluding existing BS areas and bodies of water, as, being the "future", we did not know observed transition locations in the projection period. This iterative disaggregation began with the last observed extents in year *t1* (2010) and, within each unit *i*, if we had *n* number of predicted transitions for our given projected year, we selected pixels in unit *i* with the n^{th} highest annually adjusted probabilities, and transitioned them from non-BS-to-BS. This is in line with Nieves et al. [16], Tayyebi et al. [77], Linard et al. [28] and others where it is assumed that pixels with higher transition probabilities are more likely to transition than pixels with lower probabilities. We repeated this process for all years in the projection period, using the previously projected year as the prior BS extents to expand upon, and output the union of the prior extents and the new projected transition as the next year's BS extents (Figure 1). All resulting and derived data are provided in the linked data repository (https://data.mendeley.com/datasets/cm6bnzvzfj/1).

2.3. *Analysis*

Validation and Comparison Metrics

We validated BSGMe projected extents against the withheld ESA extents for 2011, 2012, 2013, 2014, and 2015. The ESA data themselves are an imperfect reference, but our goal was to replicate the pattern of ESA's capture of BS relative to BS population and BS population density changes. Therefore, "True" in all of these validations represents agreement of the BSGMe projections with the temporally corresponding withheld ESA validation extents and "False," equally, represents disagreement. For every year, we classified every pixel in the study areas as either True Positive, False Positive, False Negative, or True Negative, TP, FP, FN, TN, respectively. Using these pixel-level designations, we calculated classification contingency-table metrics, listed in Table 3, at the unit-level.

Table 3. Classification metrics used in assessing the model performance.

Metric	Equation	Range and Interpretation
Recall (Rogan and Gladen, 1978)	$\frac{TP}{TP + FN}$	0 (no recall) – 1 (perfect recall)
Precision (Rogan and Gladen, 1978)	$\frac{TP}{TP + FP}$	0 (no precision) – 1 (perfect precision)
F1 score	$\frac{2 * \frac{TP}{TP + FP} * \frac{TP}{TP + FN}}{\frac{TP}{TP + FP} + \frac{TP}{TP + FN}}$	0 (worst) – 1 (best)

The fact that most BS land cover and non-BS land cover is in agreement from any time A to near future time B is simply due to the fact that most land cover remains the same, i.e. persistence, causes issues when looking at classification metrics [79]. Some methods exist for accounting for this [79], but because our input datasets assume that "once BS, always BS", we cannot utilise these adjustments in our binary classification assessment. Hence, the best alternative is to compare all results to a null or naïve model [79]. We utilised a conservative naïve model where we assumed that the 2010 BS extents remained constant through 2015, i.e. lacking any other information we assumed the BS extents remain approximately the same over the short-term. In end user applications, when missing year-specific BS extents, the last available BS extents are commonly used as a substitute. We validated the 2010 extents following the same procedures to compare to the modelled extents.

We also visually compared the 2015 BSGMe modelled extents and the withheld ESA 2015 extents to 2015 true colour imagery, available via Google Earth [80], to better understand areas of over/under prediction. We further carried out a quantitative classification validation of the 2015 BSGMe modelled extents and the withheld ESA RS-derived 2015 extents against the presence of OSM building footprints in Switzerland around the municipalities of Visp, Brig-Glis, Naters, and Ried-Brig, where relative model over prediction appeared to be exceptionally bad.

3. Results

Looking at the distribution of unit-level F1 scores in Figure 3, we show that all models decrease in performance as projection horizon increases, with Vietnam having the most rapid rate of decrease and largest net decrease. In all countries, it appears that the naïve model outperforms all other models to varying degrees, but not typically by much in all countries with the exception of Uganda (Figure 3).

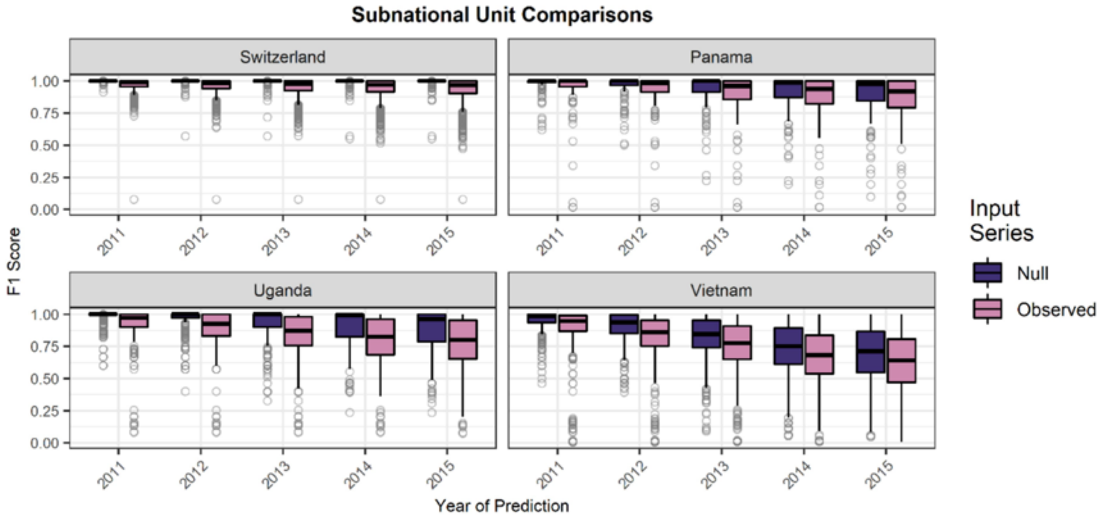


Figure 3. Boxplots of unit-level F1 scores across countries and years in the projection period and divided by the input time series to the BSGMe framework. All F1 scores were calculated by comparing pixel-level agreement/disagreement with withheld annual European Space Agency (ESA) Remote Sensing (RS)-derived extents. The median is indicated by the black line and outliers (outside of 1.5*the interquartile range) are given by grey circles.

Further investigating the distributions of F1 scores, in Figure 4, we show that recall also decreases as the projection horizon increases with Vietnam again having the most rapid and largest net decrease in recall. This makes sense as, according to the ESA RS-derived extent datasets, Vietnam had the largest relative growth while Switzerland, whose recall distributions are near identical and perfect across all input series, had very little growth, i.e. recall is driven here in Switzerland largely by persistence (Figure 4, Table 1). As expected, as the projection year increases, the recall of the BSGMe produced projections outperforms the naïve model by an increasing magnitude. Unexpectedly, considering Figure 4, Uganda had relatively high values of recall, although the variance of unit-level recall was the largest of our study countries (Figure 4).

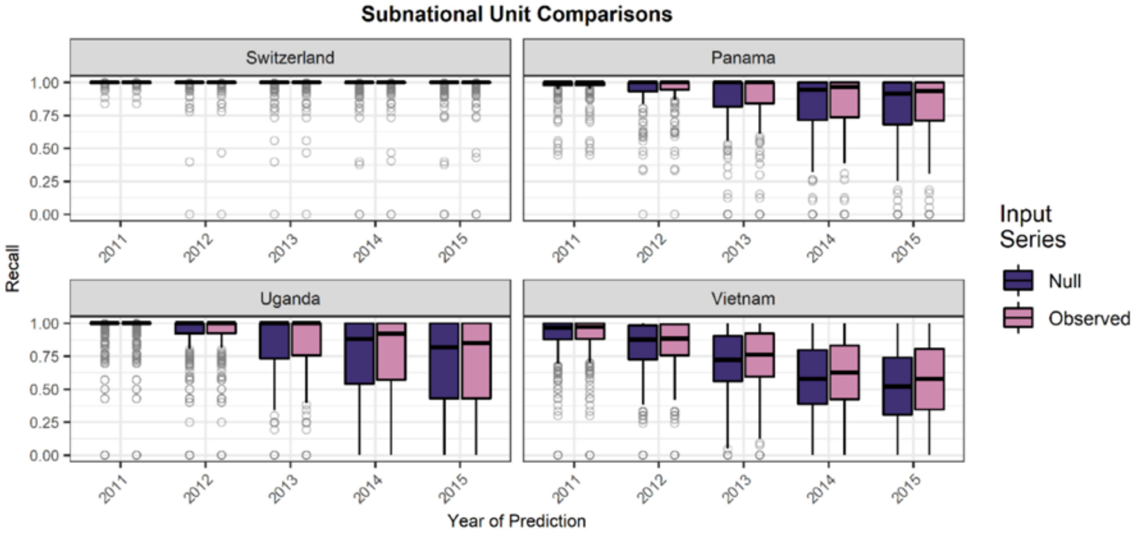


Figure 4. Boxplots of unit-level recall scores across countries and years in the projection period and divided by the input time series to the BSGMe framework. All recall values were calculated by comparing pixel-level agreement/disagreement with withheld annual ESA RS-derived extents. The median is indicated by the black line and outliers (outside of 1.5*the interquartile range) are given by grey circles.

Looking at the distribution of precision values in Figure 5, precision values decrease as the projection year increases across all countries and input series, except the naïve model because false positive could not occur with the extents remaining static. The low and variable precision shown by Uganda (Figure 5) potentially explains the observed variance of its F1 scores (Figure 3). Our best guess for the low precision here was that the ESA RS-derived extents were not as good as the population data in Uganda, i.e. leading to worse demand quantification and spatial allocation in the production of the time-series and propagating error through the BSGMe projections.

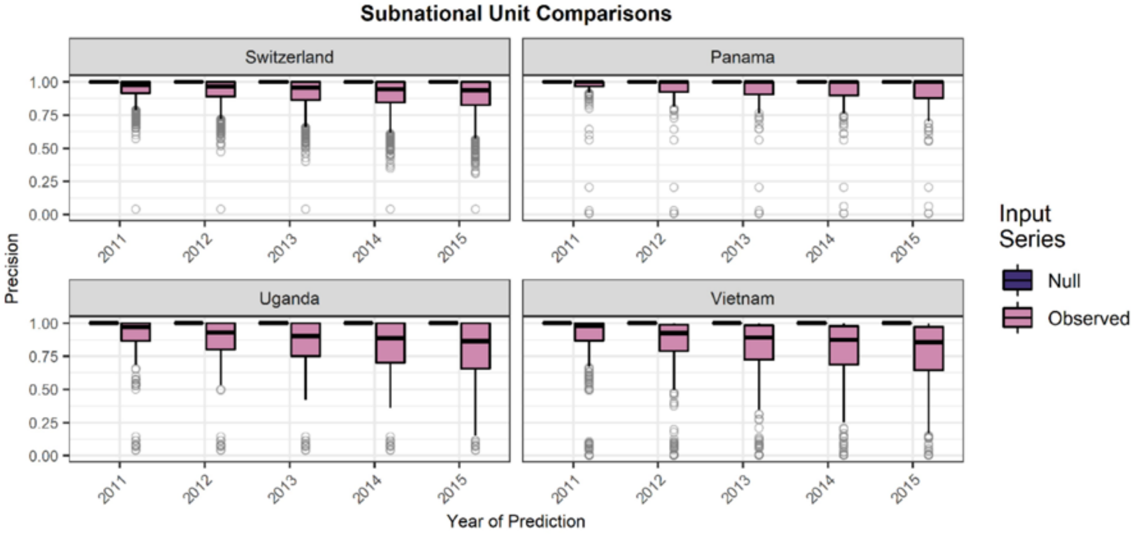


Figure 5. Boxplots of unit-level precision scores across countries and years in the projection period and divided by the input time series to the BSGMe framework. All precision values were calculated by comparing pixel-level agreement/disagreement with withheld annual ESA RS-derived extents. The median is indicated by the black line and outliers (outside of 1.5*the interquartile range) are given by grey circles.

Examining the predicted and observed extents of even a subset of projection years and areas within the study countries, Figure 6, gives some context for the findings in Figures 3–5. The same temporal trend of increases in "false positives", red in Figure 6, imply large areas of over-prediction relative to areas of agreement with ESA RS-derived extents. Areas of false negatives, blue in Figure 6, when examined against time-specific true colour imagery [80] seem to consistently coincide with low to mid-density areas of BS intermixed with trees.

Of these examples, Kampala, Uganda appeared to have the greatest magnitude of "false positives" while the Visp and Brig area of Switzerland appeared to have the largest relative number of "false positives" to RS-based observed transitions (Figure 6). This prompted us to investigate these areas more with time specific true-colour imagery [80]. Looking at time-specific true colour imagery in an area of west Kampala, Uganda, we overlaid the observed (ESA) and predicted (BSGMe) extents at 2015(Figure 7). We see that all extents are missing areas of BS, with ESA RS-derived extents missing the more fragmented and less densely settled areas (Figure 7). Within the West Kampala, Uganda scene (Figure 7, bottom left), the 2015 BSGMe-derived extents appear to have large numbers of false positives relative to the 2015 ESA RS-derived extents. Interpreting the 2015 imagery in conjunction with the extents, it is apparent that the BSGMe extents are exhibiting better recall of true BS extents (Figure 7, bottom left), suggesting that perhaps the findings of Figures 3–5 are conservative relative to the true BS extents. Although less dramatic, this was the generally the case in numerous other areas of Uganda and the other sampled countries (Figure 7, top left). However, there were examples (Figure 7, bottom right) where the BSGMe extents underestimated the true BS extents and where false positives did occur.

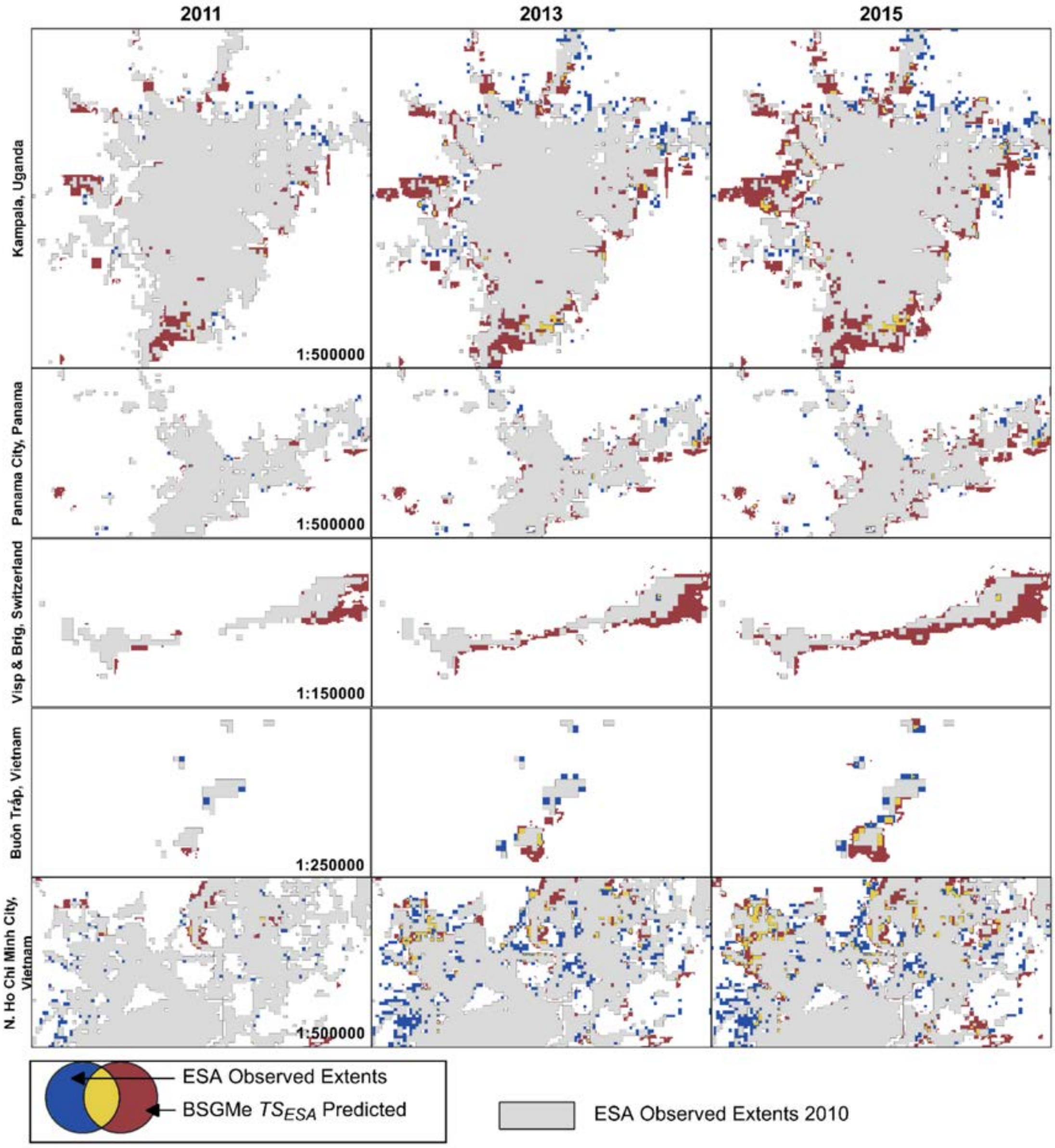


Figure 6. Map of select areas from the study countries and the projection period showing the predicted extents derived from the BSGMe (red) as well as the withheld ESA observed extents (blue). Areas where the BSGMe-derived extents and the ESA RS-derived extents agreed are shown in yellow.

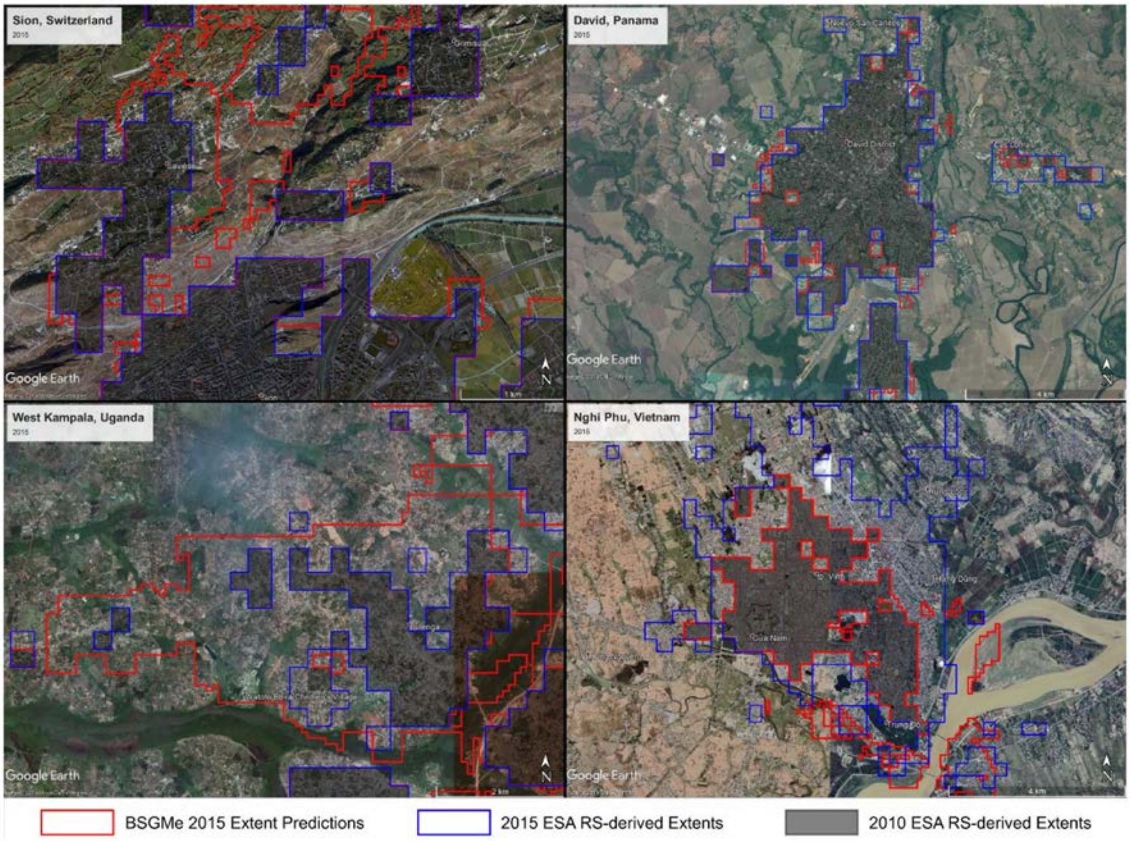


Figure 7. The 2015 BSGMe-derived extents (red), the 2015 ESA RS-derived extents (blue), and the 2010ESA RS-derived extents (transparent black areas) of BS overlain on 2015 true colour imagery via Google Earth. Map Imagery: Google, Maxar Technologies, Centre National D' Etudes Spatiales CNES/Airbus.

To begin to approach estimating how much this overestimation of false positives might be, we decided to compare an area of what appeared to be extreme over prediction by the2015 BSGMe extents relative to the 2015 ESA RS-derived extents, around Visp and Brig, Switzerland, and validate both by using the corresponding manually delineated building footprints (OpenStreetMap Contributors, 2019). By 2015, the ESA RS-derived data said there was 1,477 pixels of BS while the BSGMe-derived extents predicted 2,557 pixels of BS (Figure 8). When we compared these extents OSM building footprint data, corresponding to those present in 2015 and with near 100% coverage, across 11,966 3 arc-second pixels in the validation area, we showed that many of the areas are, in fact, not false positives (Figure 8). In fact, the observed ESA data only has a recall of 41.1% compared to the BSGMe performance of 57.9%, but the ESA extents do retain the highest precision of 84.1% (Figure 8). Considering both recall and precision simultaneously, we see that the BSGMe extents have a F1 score of 0.625 which represents approximately a 12% increase in F1 score to the ESA data (0.552) garnered by a 50% increase in recall, but at the expense of a 20% decrease in precision (Figure 8).

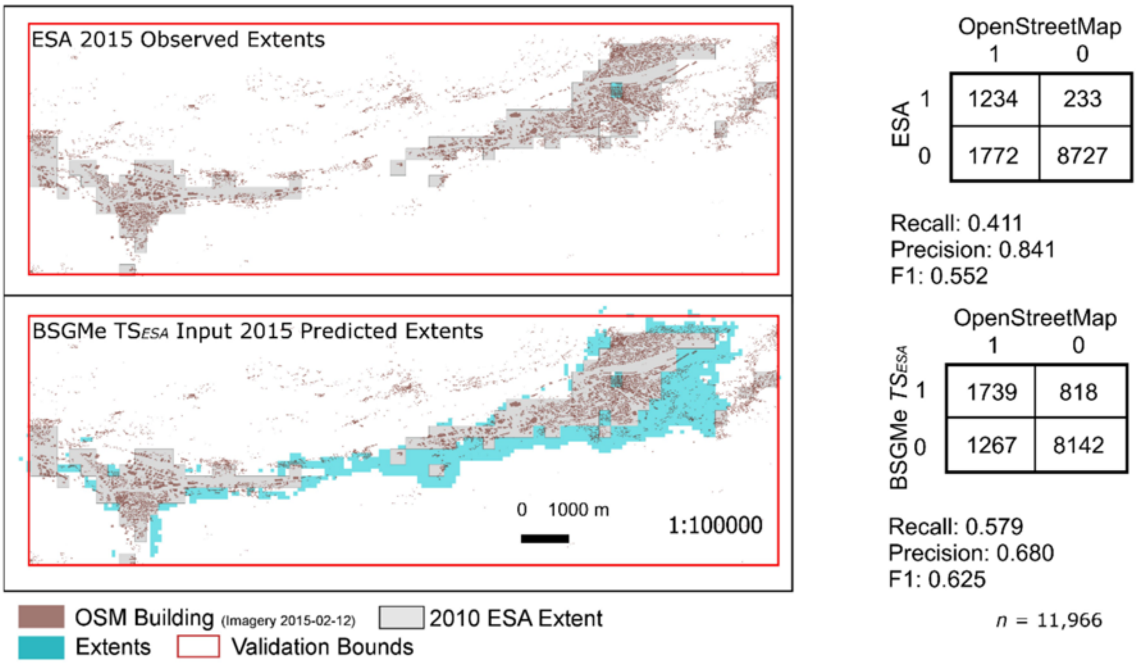


Figure 8. Validation maps of 2015 Open Street Maps (OSM) and manually delineated building footprints of the Visp and Brig area of Switzerland as compared to the ESA RS-derived extents (top left), the BSGMe TS_{ESA} predicted extents (bottom left) along with their corresponding confusion matrices and select classification metrics (right side).

4. Discussion

We have shown that the BSGMe projects BS extents into the near future with, in many cases, large agreement with the input dataset's withheld observations for predicted years (Figures 3–5). Beyond this, we found support that the validation of the BSGMe predictions, relative to the ESA RS-based observations, could be underestimating the true accuracy (Figures 7 and 8). We displayed this visually for a large proportion of Kampala, Uganda, and other example areas (Figure 7). Further, we quantified it by comparing against manually delineated building footprints for smaller settlements of the Visp and Brig area in Switzerland, showing the BSGMe having a 40 percent increase in recall and a 12 percent increase in F1 score relative to the ESA RS-based data (Figure 8).

Overall, there are inherent limits to the BSGMe approach. The framework is sensitive to the size and configuration of the subnational units used, per the Modifiable Areal Unit Problem (MAUP) [81]. We would expect that less certainty in the spatial allocation would accompany larger unit area, but the effect of unit size on demand quantification is less clear; although Nieves et al. [16] found that smaller unit size was associated with higher overall unit interpolative accuracy. Additionally, we believe that the framework would be highly sensitive to the input projected population data, yet this characteristic could have potential utility for exploring deterministic outcomes of various input urban population projection scenarios. To further clarify, while here we utilised the Stevens et al. (2015) method for producing unit level estimates of BS population, any unit-level estimates of BS population and BS population density can be provided to the BSGMe modelling framework.

Given the dasymetric nature of the BSGMe framework, measures of uncertainty that would otherwise be generated by the RF, ARIMA, ETS, and GLM models within the framework cannot be propagated to the end predicted BS extents [54,60,61,82]. This uncertainty propagation limit is similar to and was noted in the interpolative settlement modelling framework of Nieves et al. [16]. However, in general, it would be expected that the accuracy of BSGMe extrapolative predictions would have a positive relationship with any errors associated with its input datasets. For instance, if the user-selected representation of BS or estimates of BS population were relatively inaccurate, it would be logical to suppose that the framework would be tasked with sorting out noisier relationships between relative population change and BS extent expansion, and likely have poorer framework performance. In light of the framework relationships to input data error/uncertainty and the limits of propagating and quantifying this uncertainty, it is recommended that any user of this framework compare modelled outputs to the input data layers as well as the uncertainty metrics of the individual framework modelling components, which are recorded in tabular format by the framework code (see code in linked data repository https://data.mendeley.com/datasets/cm6bnzvzfj/1).

Due to persistence, the future BS projections with the highest agreement was the naïve model (Figure 3). Ignoring the actual ground truth, the model comparisons by metrics without potential end-user context are an oversimplification, with metrics like accuracy and F1 score treating a false-positive disagreement equally bad as a false-negative disagreement. It is more useful to interpret the results with a user's defined loss function in mind [83]. Should the user want to have few disagreements of any type, then the naïve model extents would be logical. However, if the user-defined cost of missing new BS extents would be a greater loss than the alternative cost of additional false-positives, the user would likely avoid the naïve approach in favor of one of the BSGMe predicted extents. Combined with the fact the false-positives of the BSGMe validations are likely inflated (Figures 3–8), it is likely that the difference in precision performance from that of the naïve model is smaller than presented here.

It is important to note that these validation findings are specific to the input ESA RS-derived extents data and the spatial scale of the input representation of BS (originally 10 arc second and then resampled to 3 arc seconds). More generally, the model framework presented here can accept any binary input of urban/built-environment/built-settlement. Although, given the framework's strong reliance on relative changes in population being indicative of relative changes in urban/built-environment/built-settlement, a functional definition of urban/built-environment/built-settlement that corresponds with aspects of the built-environment more likely to be spatially coincident with populations would be most appropriate. Whether our assumptions of population being usable as a proxy for the underlying drivers of BS expansion holds at other spatial scales of BS representation, e.g. 30m Landsat-derived, 12.5m radar-derived, or 500m Moderate Resolution Imaging Spectroradiometer (MODIS)-derived, remains unclear. Supplemental findings for a city-based area, from Nieves et al. [16], observed decreased interpolative agreement when applied to a 1 arc second radar-optical dataset, rescaled to 3 arc seconds. Theoretically, we would expect individual agency, local planning conditions, micro-economic level decision-making, and other "intangibles" from a country, to a global-extent application standpoint, to have a much larger role in the siting of BS at the average individual building scale (~ 10m–30m). However, most of this type of data, if it exists, remains unavailable across large extents and across time when working in low- to middle-income contexts.

The utilization of land cover to estimate a continuous population surface, using that population surface to estimate BS population, aggregate the BS population to the unit level to then estimate non-BS-to-BS transition demand at the unit level naturally raises a concern of circular reasoning or endogeneity. From a modelling perspective, the larger more important question is, "Why is the model being developed and what questions does it attempt to address?". Our purpose here was to develop a modelling framework capable of accurately predicting near future built-settlement expansion and to answer the question of whether this could be done by looking at subnational changes in population counts corresponding to BS areas. With this in mind, we do not believe circular reasoning, or "endogeneity", is a significant issue for the following reasons. First, our modelling framework is not an explanatory model [84] and endogeneity is, by definition, an issue of causality. Our framework falls somewhere between predictive and descriptive in nature [85] and makes no attempts at statistical inference of causation in any of its components; even the random forest is algorithmic in nature [84,86]. We were interested in utilising the correlations in our framework to create the best predictions possible, not to infer anything on the causal linkages. Secondly, there is precedent for using the hierarchical structure of population data in this manner; other model frameworks have used changes in population, at a spatial scales coarser than the scale of prediction, to quantify demand for urban area expansion [16,26–29,77,87], with one even using pixel level population to drive pixel level transitions [88]. Further, Angel et al. [3] also used geospatial and remotely sensed data to determine estimates of "urban" population and population densities that were subsequently aggregated and then used to predict future urban areas. However, this does raise the issue of fitness for purpose, similar to the discussion in Leyk et al. [11], where end users interested in causal questions and wishing to utilise datasets produced with certain covariates should assess how it was created to avoid the issue of endogeneity.

As expected, we observed that as the time from the last observation increased, the BSGMe projection decreased in agreement with the withheld ESA RS-derived validation extents. This positive association between time from last observation and projection agreement/accuracy is inherent to extrapolative models but could likely be reduced by using longer input time series, should data allow. While the automatic fitting procedure for ARIMA and ETS class models has been shown to have consistently good performance in the short-term (5–6 time steps) [64], this is predicated upon substantially longer time series (20 to 144 observations in the cited M3 competition series data [89]) than are typical with current BS or urban based population datasets at subnational unit level and with large or global extent. Due to the growing uncertainty that accompanies longer projections, we do not recommend extending this framework past the short-term without longer input time series and without further assessment. Another reason we do not currently recommend using the framework for longer term predictions is the lack of including other causal aspects of non-BS -to-BS transition, e.g. economic and planning/zoning information. We excluded such data from the framework because it is typically not available globally, for multiple time points, and at subnational resolutions.

We save unit- and year-specific 95% confidence intervals produced, via bootstrapping, by the ARIMA and ETS models [64], but we did not produce similar intervals for the GLM models (see linked data repository https://data.mendeley.com/datasets/cm6bnzvzfj/1). This was because we were only utilising the GLMs to capture the general linear trend and not inferring the true value bounds, due to an inability check for the necessary corresponding inferential assumptions for every subnational unit in an automatic, efficient, and robust manner.

5. Conclusions

Here, we have shown the BSGMe model framework to be flexible and automatable across several environmental, urban morphological and input-data quality contexts while maintaining acceptable agreement with validation data and even surpassing the performance of the input dataset's withheld observations when compared to manually interpreted conditions in time-specific true-colour imagery. This framework is novel in that it is globally applicable, with no need for user or expert input parameters, and relies largely on relative changes in subnational population to determine the timing and magnitude of changes. While validated across four countries, this framework is scalable to producing global extents across different periods and with different input BS and population datasets. Proof in point, the WorldPop Programme (www.worldpop.org) adopted this modelling framework to produce global annual BS extents at 100m resolution from 2015 through 2020, using input time-series from 2000–2014 based upon observed and BSGMi interpolated extents (https://doi.org/10.5258/SOTON/WP00649) derived from Global Human Settlement Layer, ESA urban land cover class, and Global Urban Footprint [34].

Being able to produce annual datasets of near future BS extents, and the intermediate BS populations, have a variety of end user applications where investigating potential impacts of BS population changes and BS spatial expansion can have impact, such as public health, sustainability, planning and infrastructure, and transportation management. However, as seen in this study, users should utilise auxiliary data in conjunction with their expert and or local knowledge of the application/study area to assess whether the modelled extents are suitable for their applications and needs. Additionally, this framework and its open-source code can be used as a platform for further investigating deterministic relationships between population, population densities, and BS expansion.

The extent predictions of the BSGMe framework can also be utilised, in a setup similar to this study, by producers of future BS and urban feature data sets to re-investigate areas of disagreement between the BSGMe and their extraction algorithm, knowing that there is a heightened probability of BS being truly present (Figures 7 and 8).

As the temporal resolution of global BS and urban feature data sets catch up to their high spatial resolution, further investigations of this framework will become more accessible and feasible as well as have reduced uncertainty in their conclusions. However, as evidenced in this study, there is a continued need for an independent multi-temporal data set of urban features with global extent that can be used for training and or validation. While OSM offers global extent, it has its own biases in completeness [50,51] and, more significantly, lacks any temporal attributes. One potential solution would be for the producers of urban feature data sets to make their manually identified training and validation points, footprints, and sample grid cells publicly available, e.g. by some research collaboration akin to POPGRID Data Collaborative (www.popgrid.org) with agreed upon documentation, data attribute, and definitional standards. Until such a time, large scale ground truthing, much less temporal ground truthing, of BS or urban features will likely be limited and often surpass the resources of many studies with large or global extent.

Future work should investigate the robustness of this framework with different spatial scale representations of BS as inputs and differing lengths of input time-series. Additional experimentation with the demand forecasting methods is also a large area that remains to be explored. Further validation of more areas should also be prioritized, particularly in areas where urban feature datasets are known to have extraction issues, e.g. arid regions, in order to understand how such error may propagate through this framework into the resulting extents. Other desirable work would involve examination of the applied utility of the BS outputs produced by both the interpolative and extrapolative BSGM frameworks.

Author Contributions: Conceptualization: J.J.N., A.S., F.R.S., A.E.G., J.E.S., and A.J.T. Data Curation: J.J.N. Formal

Analysis: J.J.N. Funding Acquisition: J.J.N., A.S., and A.J.T. Investigation: J.J.N. Methodology: J.J.N. and A.C. Project Administration: J.J.N. Resources: J.J.N., M.B., A.S., and A.J.T. Software: J.J.N., M.B., and D.K. Supervision: A.S., J.E.S., and A.J.T. Validation: J.J.N. Visualization: J.J.N. Writing—original draft: J.J.N. Writing—review and editing: All authors. All authors have read and agreed to the published version of the manuscript.

Funding: J.J.N. is funded through the Economic and Social Research Council's Doctoral Training Program, specifically under the South Coast branch (ESRC SC DTP). AS is supported by funding from the Bill & Melinda Gates Foundation (OPP1134076).

Acknowledgments: Many of the spatial covariates (doi:10.5258/SOTON/WP00644) used here are the product of the "Global High Resolution Population Denominators Project" funded by the Bill and Melinda Gates Foundation (OPP1134076). The authors acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of the data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A

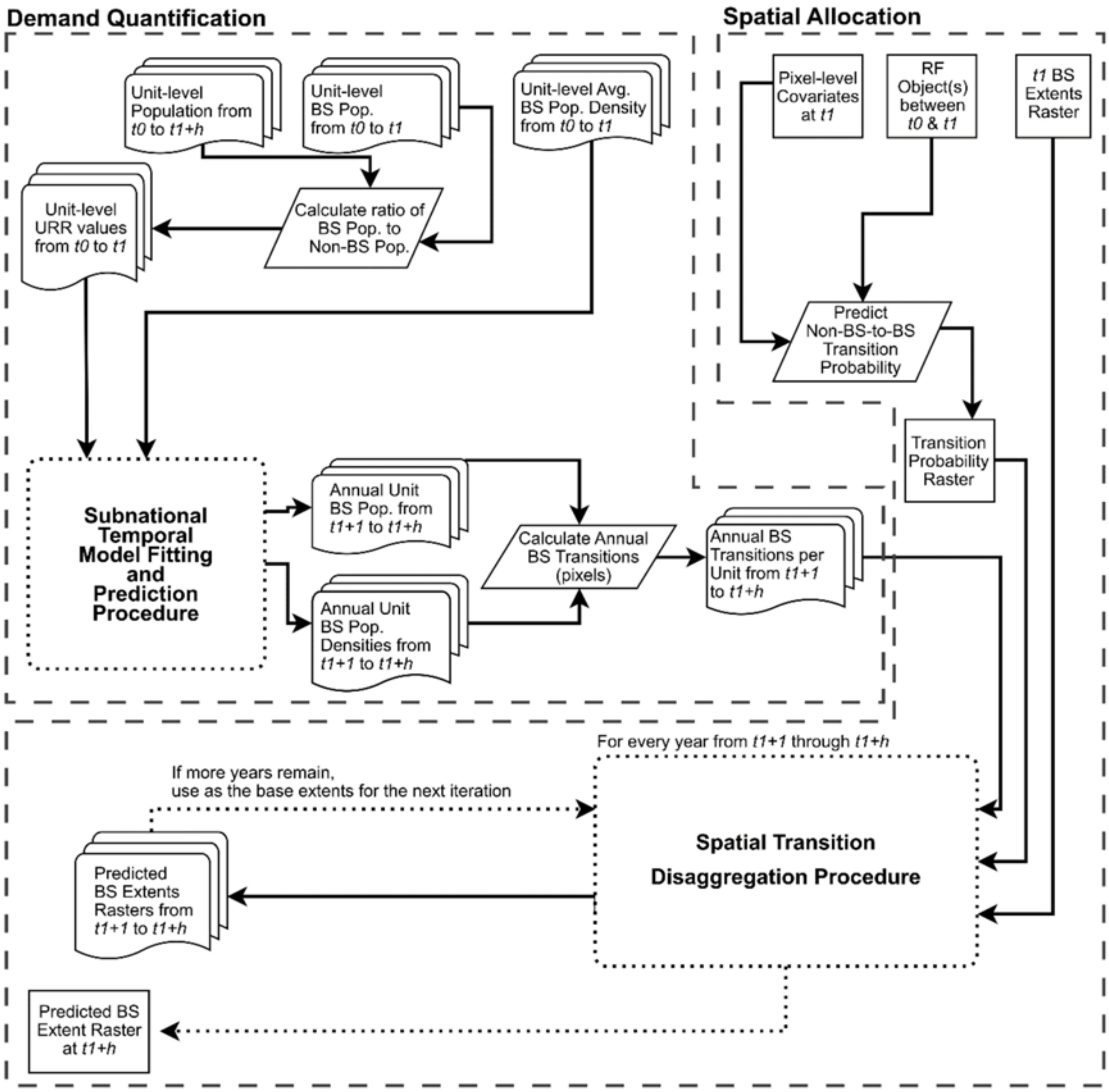

Demand Quantification
Spatial Allocation
Unit-level Population from t0 to t1+h
Unit-level BS Pop. from t0 to t1
Unit-level Avg. BS Pop. Density from t0 to t1
Unit-level URR values from t0 to t1
Calculate ratio of BS Pop. to Non-BS Pop.
Pixel-level Covariates at t1
RF Object(s) between t0 & t1
t1 BS Extents Raster
Predict Non-BS-to-BS Transition Probability
Transition Probability Raster
Subnational Temporal Model Fitting and Prediction Procedure
Annual Unit BS Pop. from t1+1 to t1+h
Annual Unit BS Pop. Densities from t1+1 to t1+h
Calculate Annual BS Transitions (pixels)
Annual BS Transitions per Unit from t1+1 to t1+h
For every year from t1+1 through t1+h
If more years remain, use as the base extents for the next iteration
Spatial Transition Disaggregation Procedure
Predicted BS Extents Rasters from t1+1 to t1+h
Predicted BS Extent Raster at t1+h

Figure A1. Full process diagram for the Built-Settlement Growth Model—extrapolation (BSGMe) as broken down into the "Demand Quantification" procedure and the "Spatial Allocation Procedure". For details on the "Spatial Transition Disaggregation Procedure", readers are referred to Nieves et al. [16]. For details on the "Subnational Temporal Model Fitting and Prediction Procedure", readers are referred to Appendix A, Figure A2.

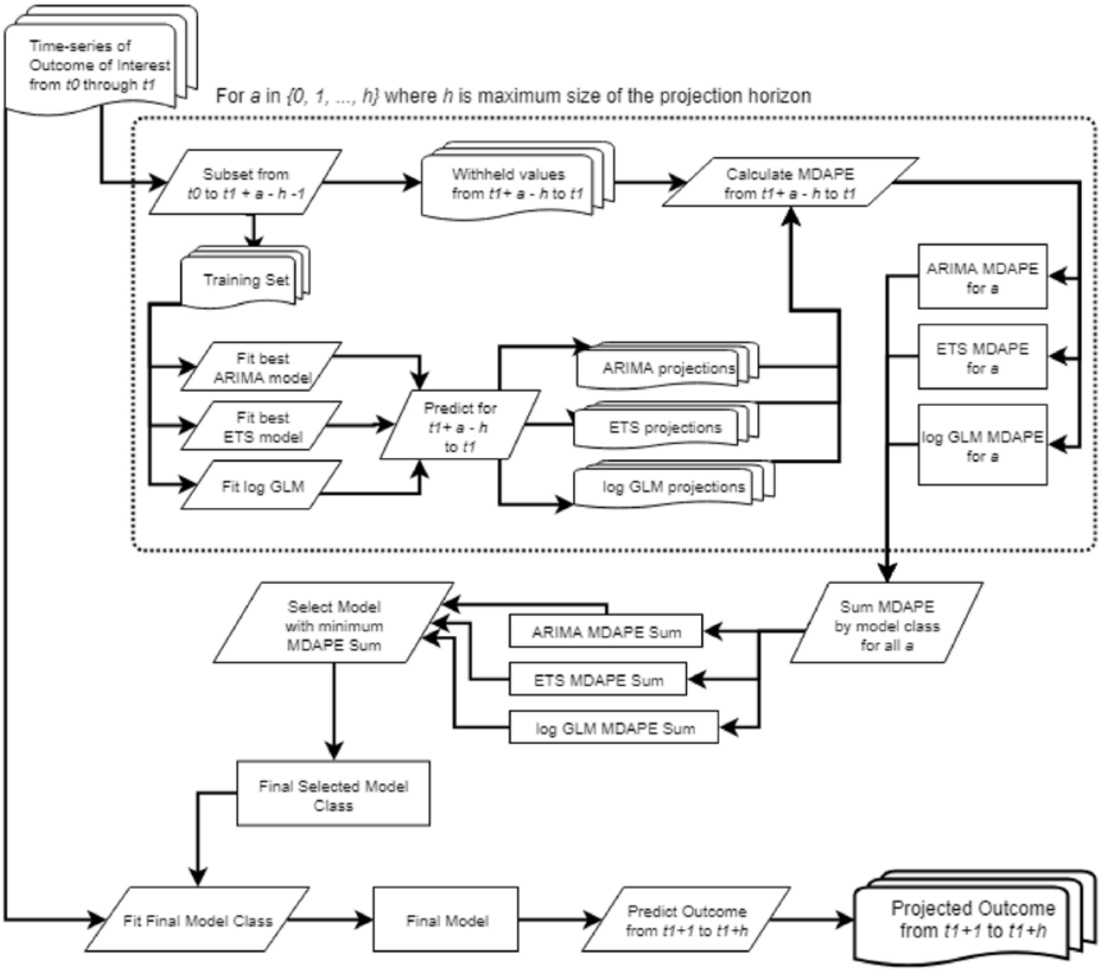

Time-series of Outcome of Interest from t0 through t1
For a in {0, 1, ..., h} where h is maximum size of the projection horizon
Subset from t0 to t1 + a - h -1
Withheld values from t1+ a - h to t1
Calculate MDAPE from t1+ a - h to t1
Training Set
ARIMA MDAPE for a
ETS MDAPE for a
log GLM MDAPE for a
Fit best ARIMA model
Fit best ETS model
Fit log GLM
Predict for t1+ a - h to t1
ARIMA projections
ETS projections
log GLM projections
Select Model with minimum MDAPE Sum
ARIMA MDAPE Sum
ETS MDAPE Sum
log GLM MDAPE Sum
Sum MDAPE by model class for all a
Final Selected Model Class
Fit Final Model Class
Final Model
Predict Outcome from t1+1 to t1+h
Projected Outcome from t1+1 to t1+h

Figure A2. Full process diagram of the "Subnational Temporal Model Fitting and Prediction Procedure" referenced in Appendix A, Figure A1. Readers are directed to the main text for acronym references and details on the rolling origin framework.

Table A1. Table of time specific, or assumed temporally invariant, covariates used in the modelling of the population surfaces following the procedure from Stevens et al. [55].

Covariate	Time Point(s)[a]	Original Source	Source Resolution
DTE Cultivated landcover	2000–2010	ESA CCI Landcover [36] classes 10–30	10 arc seconds
DTE Woody, Herbaceous, Shrub landcover	2000–2010	ESA CCI Landcover [36] classes 40–120	10 arc seconds
DTE Grassland landcover	2000–2010	ESA CCI Landcover [36] class 130	10 arc seconds
DTE Lichens and Mosses landcover	2000–2010	ESA CCI Landcover [36] class 140	10 arc seconds
DTE Sparse Vegetation landcover	2000–2010	ESA CCI Landcover [36] classes 150–153	10 arc seconds
DTE Aquatic Vegetation landcover	2000–2010	ESA CCI Landcover [36] classes 160–180	10 arc seconds
DTE Bare Areas	2000–2010	ESA CCI Landcover [36] class 200	10 arc seconds

Table A1. *Cont.*

Covariate	Time Point(s)[a]	Original Source	Source Resolution
DTE Built-settlement	2000–2010	ESA CCI Landcover [36] class 190	
Distance to Inland Water Bodies	2015, assumed invariant	MERIS-based water bodies [39]	5 arc seconds
Distance to Roads	Downloaded 2017, assumed invariant as temporally specific road data unavailable	OpenStreetMap [44]	Vector
Distance to Rivers	Downloaded 2017, assumed invariant	OpenStreetMap [44]	Vector
Distance to Coastline	Based upon boundaries of GPWv4, assumed invariant	CIESIN GPWv4 [40]	Vector
Slope	2000, assumed invariant	World Wildlife Fund Void-filled Hydrosheds [37]	3 arc seconds
Elevation	2000, assumed invariant	World Wildlife Fund Void-filled Hydrosheds [37]	3 arc seconds

DTE: Distance To nearest Edge a Note, for any covariate derived from land cover or built-settlement, only one year-specific covariate was used corresponding to the desired population surface (e.g., for a 2000 population surface only covariates corresponding to 2000, or those assumed temporally invariant, were used as covariates).

References

1. United Nations. *World Urbanization Prospects: The 2018 Revision*; United Nations: New York, NY, USA, 2018.
2. Ledent, J. Rural-Urban Migration, Urbanization, and Economic Development. *Econ. Dev. Cult. Change* **1982**, *30*, 507–538. [CrossRef]
3. Angel, S.; Parent, J.; Civco, D.L.; Blei, A.M.; Potere, D. The Dimensions of Global Urban Expansion: Estimates and Projections for All Countries, 2000-2050. *Prog. Plann.* **2011**, *75*, 53–107. [CrossRef]
4. Cohen, B. Urban growth in developing countries: A review of current trends and a caution regarding existing forecasting. *World Dev.* **2004**, *32*, 23–51. [CrossRef]
5. Espey, J. Sustainable development will falter without data. *Nature* **2019**, *571*, 299. [CrossRef] [PubMed]
6. Solecki, W.; Seto, K.C.; Marcotullio, P.J. It's Time for an Urbanization Science. *Environ. Sci. Policy Sustain. Dev.* **2013**, *55*, 12–17. [CrossRef]
7. Scott, G.; Rajabifard, A. Sustainable Development and Geospatial Information: A Strategic Framework for Integrating a Global Policy Agenda into National Geospatial Capabilities. *Geo-spatial Inf. Sci.* **2017**, *20*, 59–76. [CrossRef]
8. United Nations. *United Nations Transforming Our World: The 2030 Agenda for Sustainable Development*; United Nations: New York, NY, USA, 2016.
9. United Nations. *Economic and Social Council Report of the High-Level Political Forum on Sustainable Development Convened under the Auspices of the Economic and Social Council at its 2016 Session*; United Nations: New York, NY, USA, 2016.
10. Freire, S.; Schiavina, M.; Florczyk, A.J.; MacManus, K.; Pesaresi, M.; Corbane, C.; Borkovska, O.; Mills, J.; Pistolesi, L.; Squires, J.; et al. Enhanced data and methods for improving open and free global population grids: putting 'leaving no one behind' into practice. *Int. J. Digit. Earth* **2018**, 1–17. [CrossRef]
11. Leyk, S.; Gaughan, A.E.; Adamo, S.B.; de Sherbinin, A.; Balk, D.; Freire, S.; Rose, A.; Stevens, F.R.; Blankespoor, B.; Frye, C.; et al. The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth Syst. Sci. Data* **2019**, *11*, 1385–1409. [CrossRef]
12. Pesaresi, M.; Guo, H.; Blaes, X.; Ehrlich, D.; Ferri, S.; Gueguen, L.; Halkia, S.; Kauffmann, M.; Kemper, T.; Lu, 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*, 2102–2131. [CrossRef]
13. Esch, T.; Marconcini, M.; Felbier, A.; Roth, A.; Heldens, W.; Huber, M.; Schwinger, M.; Taubenbock, H.;Muller, A.; Dech, S. 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*, 1617–1621. [CrossRef]
14. Esch, T.; Bachofer, F.; Heldens, W.; Hirner, A.; Marconcini, M.; Palacios-Lopez, D.; Roth, A.; Üreyen, S.; Zeidler, J.; Dech, S.; et al. Where We Live—A Summary of the Achievements and Planned Evolution of the Global Urban Footprint. *Remote Sens.* **2018**, *10*, 895. [CrossRef]
15. Corbane, C.; Pesaresi, M.; Politis, P.; Syrris, V.; Florczyk, A.J.; Soille, P.; Maffenini, L.; Burger, A.; Vasilev, V.; Rodriguez, D.; et al. Big earth data analytics on Sentinel-1 and Landsat imagery in support to global human settlements mapping. *Big Earth Data* **2017**, *1*, 118–144. [CrossRef]
16. Nieves, J.J.; Sorichetta, A.; Linard, C.; Bondarenko, M.; Steele, J.E.; Stevens, F.R.; Gaughan, A.E.; Carioli, A.; Clarke, D.J.; Esch, T.; et al. Annually modelling built-settlements between remotely-sensed observations using relative changes in subnational populations and lights at night. *Comput. Environ. Urban Syst.* **2020**, *80*, 101444. [CrossRef] [PubMed]
17. Florczyk, A.J.; Melchiorri, M.; Zeidler, J.; Corbane, C.; Schiavina, M.; Freire, S.; Sabo, F.; Politis, P.; Esch, T.; Pesaresi, M. The Generalised Settlement Area: mapping the Earth surface in the vicinity of built-up areas. *Int. J. Digit. Earth* **2019**, 1–16. [CrossRef]
18. Pesaresi, M.; Ehrlich, D.; Ferri, S.; Florczyk, A.J.; Freire, S.; Halkia, S.; Julea, A.M.; Kemper, T.; Soille, P.; Syrris, V. *Operating Procedure for the Production of the Global Human Settlement Layer from Landsat*

Data of the Epochs 1975, 1990, 2000, and 2014; Publications Office of the European Union: Brussels, Belgium, 2016.

19. ESA; CCI. *European Space Agency Climate Change Initiative Landcover*; ESA: Paris, France, 2016.
20. Facebook Connectivity Lab; Columbia University Center for International Earth Science Information Network (CIESIN). *High Resolution Settlement Layer*; CIESIN: Palisades, NY, USA, 2016.
21. Small, C.; Cohen, J.E. Continental physiography, climate, and the global distribution of human population. *Curr. Anthropol.* **2004**, *45*, 269–277. [CrossRef]
22. Small, C.; Elvidge, C.D.; Balk, D.; Montgomery, M. Spatial scaling of stable night lights. *Remote Sens. Environ.* **2011**, *115*, 269–280. [CrossRef]
23. Linard, C.; Gilbert, M.; Snow, R.W.; Noor, A.M.; Tatem, A.J. Population Distribution, Settlement Patterns and Accessibility across Africa in 2010. *PLoS One* **2012**, *7*, e31743. [CrossRef]
24. Seto, K.C.; Fragkias, M.; Guneralp, B.; Reilly, M.K. A Meta-Analysis of Global Urban Land Expansion. *PLoS ONE* **2011**, *6*, e23777. [CrossRef]
25. Batty, M. Urban Modeling. In *International Encyclopedia of Human Geography*; Elsevier: Oxford, UK, 2009; pp. 51–58.
26. Sante, I.; Garcia, A.M.; Miranda, D.; Crecente, R. Cellular Automata Models for the Simulation of Real-world Urban Processes: A Review and Analysis. *Landsc. Urban Plan.* **2010**, *96*, 108–122. [CrossRef]
27. Li, X.; Gong, P. Urban growth models: progress and perspective. *Sci. Bull.* **2016**, *61*, 1637–1650. [CrossRef]
28. Linard, C.; Tatem, A.J.; Gilbert, M. Modelling Spatial Patterns of Urban Growth in Africa. *Appl. Geogr.* **2013**, *44*, 23–32. [CrossRef] [PubMed]
29. Seto, K.C.; Guneralp, B.; Hutyra, L.R. Global Forecasts of Urban Expansion to 2030 and Direct Impacts on Biodiversity and Carbon Pools. *Proc. Natl. Acad. Sci. USA* **2012**, *109*, 16083–16088. [CrossRef] [PubMed]
30. Schneider, A.; Mertes, C.M.; Tatem, A.J.; Tan, B.; Sulla-Menashe, D.; Graves, S.J.; Patel, N.N.; Horton, J.A.; Gaughan, A.E.; Rollo, J.T.; et al. A new urban landscape in East–Southeast Asia, 2000–2010. *Environ. Res. Lett.* **2015**, *10*. [CrossRef]
31. Goldewijk, K.K.; 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*, 565–573. [CrossRef]
32. de Koning, G.H.J.; Verburg, P.H.; Veldkamp, A.; Fresco, L.O. Multi-scale modelling of land use change dynamics in Ecuador. *Agrcultural Syst.* **1999**, *61*, 77–93. [CrossRef]
33. Verburg, P.H.; Soepboer, W.; Veldkamp, A.; Limpiada, R.; Espladon, V.; Mastura, S.S.A. Modeling the Spatial Dynamics of Regional Land Use: The CLUE-S Model. *Environ. Manage.* **2002**, *30*, 391–405. [CrossRef] [PubMed]
34. Lloyd, C.T.; Chamberlain, H.; Kerr, D.; Yetman, G.; Pistolesi, L.; Stevens, F.R.; Gaughan, A.E.; Nieves, J.J.; Hornby, G.; MacManus, K.; et al. Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets. *Big Earth Data* **2019**, *3*, 108–139. [CrossRef]
35. Tobler, W.; Deichmann, U.; Gottsegen, J.; Maloy, K. World Population in a Grid of Spherical Quadrilaterals. *Int. J. Popul. Geogr.* **1997**, *3*, 203–225. [CrossRef]
36. ESA; CCI. *European Space Agency Climate Change Initiative Landcover*; ESA: Paris, France, 2017.
37. Lehner, B.; Verdin, K.; Jarvis, A. New Global Hydrography Derived from Spaceborne Elevation Data. *Eos, Trans. Am. Geophys. Union* **2008**, *89*, 93–94. [CrossRef]
38. U.N. Enviroment Programme World Conservation Monitoring Centre; IUCN World Commission on Protected Areas. *World Database on Protected Areas*; United Nations: New York, NY, USA, 2015.
39. Lamarche, C.; Santoro, M.; Bontemps, S.; D’Andrimont, R.; Radoux, J.; Giustarini, L.; Brockmann, C.; Wevers, J.; Defourny, P.; Arino, O. 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.* **2017**, *9*. [CrossRef]
40. Doxsey-Whitfield, E.; MacManus, K.; Adamo, S.B.; Pistolesi, L.; Squires, J.; Borkovska, O.; Baptista, S.R. 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*, 226–234. [CrossRef]

41. Zhang, Q.; Seto, K.C. Mapping urbanization dynamics at regional and global scales using multi-temporal DMSP/OLS nighttime light data. *Remote Sens. Environ.* **2011**, *115*, 2320–2329. [CrossRef]
42. Earth Observation Group NOAA. *VIIRS Nighttime Lights - One Month Composites*; National Centers for Environmental Information: Asheville, NC, USA, 2016.
43. Nelson, A. *Estimated Travel Time to the Nearest city of 50,000 or More People in Year 2000*; Global Environment Monitoring Unit - Joint Research Centre of the European Commission: Ispra, Italy, 2008.
44. OpenStreetMap. Contributers OpenStreetMap (OSM) Database. 2017. Available online: https://www. openstreetmap.org/ (accessed on 12 May 2020).
45. Hijmans, R.J.; Cameron, S.E.; Parra, J.L.; Jones, P.G.; Jarvis, A. Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* **2005**, *25*, 1965–1978. [CrossRef]
46. ESA CCI New Release of the C3S Global Land Cover products for 2016, 2017 and 2018 consistent with the CCI 1992 – 2015 map series. Available online: https://www.esa-landcover-cci.org/?q=node/197 (accessed on 14 November 2019).
47. UCL. *Geomatics Land Cover CCI Product User Guide Version 2.0*; UCL: London, UK, 2017.
48. Goodchild, M.F. Citizens as sensors: the world of volunteered geography. *GeoJournal* **2007**, *69*, 211–221. [CrossRef]
49. Haklay, M. How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets. *Environ. Plan. B Urban Anal. City Sci.* **2010**, *37*, 682–703. [CrossRef]
50. Neis, P.; Zipf, A. Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap. *ISPRS Int. J. Geo-Information* **2012**, *1*, 146–165. [CrossRef]
51. Fan, H.; Zipf, A.; Fu, Q.; Neis, P. Quality assessment for building footprints data on OpenStreetMap. *Int. J. Geogr. Inf. Sci.* **2014**, *28*, 700–719. [CrossRef]
52. Senaratne, H.; Mobasheri, A.; Ali, A.L.; Capineri, C.; Haklay, M. A review of volunteered geographic information quality assessment methods. *Int. J. Geogr. Inf. Sci.* **2017**, *31*, 139–167. [CrossRef]
53. Linard, C.; Tatem, A.J.; Stevens, F.R.; Gaughan, A.E.; Patel, N.N.; Huang, Z. Use of active and passive VGI data for population distribution modelling: experience from the WorldPop project. In Proceedings of the Eighth International Conference on Geographic Information Science, Vienna, Austria, 24–26 September 2014; pp. 1–16.
54. Stevens, F.R.; Gaughan, A.E.; Linard, C.; Tatem, A.J. Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-sensed Data and Ancillary Data. *PLoS One* **2015**, *10*, e0107042. [CrossRef]
55. Forget, Y.; Linard, C.; Gilbert, M. Supervised Classification of Built-Up Areas in Sub-Saharan African Cities Using Landsat Imagery and OpenStreetMap. *Remote Sens.* **2018**, *10*, 1145. [CrossRef]
56. Grippa, T.; Georganos, S.; Zarougui, S.; Bognounou, P.; Diboulo, E.; Forget, Y.; Lennert, M.; Vanhuysse, S.; Mboga, N.; Wolff, E. Mapping Urban Land Use at Street Block Level Using OpenStreetMap, Remote Sensing Data, and Spatial Metrics. *ISPRS Int. J. Geo-Information* **2018**, *7*, 246. [CrossRef]
57. Weiss, D.J.; Nelson, A.; Gibson, H.S.; Temperley, W.; Peedell, S.; Lieber, A.; Hancher, M.; Poyart, E.; Belchior, S.; Fullman, N.; et al. A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature* **2018**, *553*, 333–336. [CrossRef] [PubMed]
58. Switzerland Federal Statistical Office STAT-TAB - interaktive Tabellen. Available online: https://www.pxweb. bfs.admin.ch (accessed on 16 August 2019).
59. R Core Team. *R: A Language and Environment Layer for Statistical Computing*; R Core Team: Vienna, Austria, 2016.
60. Mennis, J.; Hultgren, T. Intelligent dasymetric mapping and its application to areal interpolation. *Cartogr. Geogr. Inf. Sci.* **2006**, *33*, 179–194. [CrossRef]
61. Mennis, J. Generating surface models of population using dasymetric mapping. *Prof. Geogr.* **2003**, *55*, 31–42.
62. Gaughan, A.E.; Stevens, F.R.; Huang, Z.; Nieves, J.J.; Sorichetta, A.; Lai, S.; Ye, X.; Linard, C.; Hornby, G.M.; Hay, S.I.; et al. Spatiotemporal patterns of population in mainland China, 1990 to 2010. *Sci. Data* **2016**, *3*. [CrossRef] [PubMed]

63. Box, G.E.P.; Jenkins, G.M. *Time Series Analysis: Forecasting and Control*, 2nd ed.; Wiley: San Francisco, CA, USA, 1976.

64. Hyndman, R.J.; Khandakar, Y. Automatic Time Series Forecasting: The forecast package for R. *J. Stat. Softw.* **2008**, *27*. [CrossRef]

65. Hyndman, R.J.; Koehler, A.B.; Snyder, R.D.; Grose, S. A state space framework for automatic forecasting using exponential smoothing methods. *Int. J. Forecast.* **2002**, *18*, 439–454. [CrossRef]

66. Pegels, C.C. Exponential Forecasting: Some New Variations. *Manage. Sci.* **1969**, *15*, 311–315.

67. Ord, J.K.; Koehler, A.B.; Snyder, R.D. Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models. *J. Am. Stat. Assoc.* **1997**, *92*. [CrossRef]

68. Hyndman, R.J.; Booth, H. Stochastic population forecasts using functional data models for mortality, fertility and migration. *Int. J. Forecast.* **2008**, *24*, 323–342. [CrossRef]

69. Fildes, R.; Petropoulos, F. Simple versus complex selection rules for forecasting many time series. *J. Bus. Res.* **2015**, *68*, 1692–1703. [CrossRef]

70. Nelder, J.A.; Wedderburn, R.W.M. Generalized Linear Models. *J. R. Stat. Soc. Ser. A* **1972**, *135*, 370–384. [CrossRef]

71. Shang, H.L. Mortality and life expectancy forecasting for a group of populations in developed countries: A multilevel functional data method. *Ann. Appl. Stat.* **2016**, *10*, 1639–1672. [CrossRef]

72. Tashman, L.J. Out-of-sample tests of forecasting accuracy: an analysis and review. *Int. J. Forecast.* **2000**, *16*, 437–450. [CrossRef]

73. Hyndman, R.J.; Booth, H.; Yasmeen, F. Coherent Mortality Forecasting: The Product-Ratio Method With Functional Time Series Models. *Demography* **2013**, *50*, 261–283. [CrossRef]

74. Makridakis, S.; Hibon, M. The M3-Competition: results, conclusions, and implications. *Int. J. Forecast.* **2000**, *16*, 451–476. [CrossRef]

75. Breiman, L. Random Forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]

76. Kamusoko, C.; Gamba, J. Simulating Urban Growth Using a Random Forest-Cellular Automata (RF-CA) Model. *ISPRS Int. J. Geo-Information* **2015**, *4*, 447–470. [CrossRef]

77. Tayyebi, A.; Pekin, B.K.; Pijanowski, B.C.; Plourde, J.D.; Doucette, J.S.; 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*, 422–442. [CrossRef]

78. Rogan, W.J.; Gladen, B. Estimating prevalence from the results of a screening test. *Am. J. Epidemiol.* **1978**, *107*, 71–76. [CrossRef]

79. Pontius, R.G.; Shusas, E.; McEachern, M. Detecting important categorical land changes while accounting for persistence. *Agric. Ecosyst. Environ.* **2004**, *101*, 251–268. [CrossRef]

80. Google Earth; Maxar Technologies; CNES/Airbus Map Imagery. 2019. Available online: https://earth.google.com/web/ (accessed on 12 May 2020).

81. Openshaw, S. The modifiable areal unit problem. *Concepts Tech. Mod. Geogr.* **1984**, *38*.

82. Nagle, N.N.; Buttenfield, B.P.; Leyk, S.; Spielman, S. Dasymetric Modeling and Uncertainty. *Ann. Assoc. Am. Geogr.* **2014**, *104*, 80–94. [CrossRef] [PubMed]

83. Savage, L.J. The Theory of Statistical Decision. *J. Am. Stat. Assoc.* **1951**, *46*, 55–67. [CrossRef]

84. Breiman, L. Statistical Modeling: The Two Cultures. *Stat. Sci.* **2001**, *16*, 199–231. [CrossRef]

85. Shmueli, G. To Explain or Predict. *Stat. Sci.* **2010**, *25*, 289–310. [CrossRef]

86. Liaw, A.; Wiener, M. Classification and Regression by randomForest. *R News* **2002**, *3*, 18–22.

87. Verburg, P.H.; Overmars, K.P. Combining top-down and bottom-up dynamics in land use modeling: exploring the future of abandoned farmlands in Europe with the Dyna-CLUE model. *Landsc. Ecol.* **2009**, *24*, 1167–1181. [CrossRef]

88. Schaldach, R.; Alcamo, J.; Koch, J.; Kölking, C.; Lapola, D.M.; Schüngel, J.; Priess, J.A. An integrated approach to modelling land-use change on continental and global scales. *Environ. Model. Softw.* **2011**, *26*, 1041–1051. [CrossRef]

89. International Institute of Forecasters M-3 Competition. Available online: https://forecasters.org/resources/ time-series-data/m3-competition/ (accessed on 1 December 2019).

Chapter 5 Measuring the contribution of built-settlement data to global population mapping

Nieves, J. J., Bondarenko, M., Kerr, D., Ves, N., Yetman, G., Sinha, P., Clarke, D. J., Sorichetta, A., Stevens, F. R., Gaughan, A. E., & A. J. Tatem. Measuring the contribution of built-settlement data to global population mapping

The version included in this thesis is the author's version of a work that was submitted for publication in *Social Sciences and Humanities Open*. Changes resulting from the publishing process, such as peer review, editing, corrections, structural formatting, and other quality control mechanisms may not be reflected in this document. Changes may have been made to this work since it was submitted for publication.

ABSTRACT

Top-down population modelling has gained applied prominence in public health, planning, and sustainability applications at the global scale. These top-down population modelling methods often rely on remote-sensing (RS) derived representation of the built-environment and settlements as key predictive covariates. While these RS-derived data, which are global in extent, have become more advanced and more available, gaps in spatial and temporal coverage remain. Here we have modelled built-settlement extents between 2000 and 2012 and demonstrate the applied utility and information provided by these annually modelled data for the application of annually modelling population across 172 countries. We demonstrate that the modelled built-settlement data are consistently the 2nd most important covariate in predicting population density, behind annual lights at night, across the globe and across the study period. Further, we demonstrate that this modelled built-settlement data often provides more information than current annually available RS-derived data and last observed built-settlement extents.

Keywords:

Urban; population; growth model; built; settlement; machine learning; meta-analysis

1. Introduction

It is projected that, by 2050, an additional 13 percent of the world's population will live in urbanized areas, with most of this growth occurring in low- to middle-income countries (Angel *et al.*, 2011; United Nations, 2018). Further, much of this projected growth will not occur in the largest cities, but rather it will occur in small to medium sized settlements (Cohen, 2004). This projected growth logically has implications for sustainable development. This has been noted in the 2030 Sustainable Development Goals (SDGs), particularly in SDG 11 for "Sustainable Communities and Cities" (United Nations, 2016). Further, the SDGs aim to make sure "no one is left behind" (United Nations - Economic and Social Council, 2016), which applies to traditionally underrepresented, overlooked, or excluded persons. This includes those small settlements that have been often missed in various measures and counts including censuses (Tatem *et al.*, 2007; Leyk *et al.*, 2019) and remotely-sensing (RS)-derived representations of settlements (BS) (Pesaresi *et al.*, 2013; Kuffer, Barros and Sliuzas, 2014; Kuffer, Pfeffer and Sliuzas, 2016; Weber *et al.*, 2018; Nieves *et al.*, 2020a). More generally, a key SDG goal is to expand the availability and accessibility of base data to help facilitate the planning, implementation, and assessment of programs to achieve the 2030 SDGs (United Nations, 2016; Scott and Rajabifard, 2017).

Since 2010, a new type of global consistent datasets of RS-derived representations of built-settlement (BS), defined as above ground structures that can support human habitation and related economic phenomena (Pesaresi *et al.*, 2013; Florczyk *et al.*, 2019; Nieves *et al.*, 2020a), have become available at single and multiple time points (Pesaresi *et al.*, 2013, 2016; Esch *et al.*, 2013, 2018; Facebook Connectivity Lab and Columbia University Center for International Earth Science Information Network - CIESIN, 2016; Corbane *et al.*, 2017; Microsoft, 2018). These data, which are better able to capture small settlements due to having spatial resolutions typically ranging from the representation of individual buildings to 40m, have been found to be highly important in top-down population modelling applications (Patel *et al.*, 2015; Nieves *et al.*, 2017; Reed *et al.*, 2018; Leyk *et al.*, 2019; Stevens *et al.*, 2020). However, this new generation of BS datasets is not without limits. Like most RS-derived products, they are limited by the quality and availability of imagery, training and validation data, atmospheric conditions, and sensor/platform errors (Pesaresi *et al.*, 2013, 2016; Esch *et al.*, 2013, 2018; Corbane *et al.*, 2017). While these new datasets have leveraged advances in imagery availability, computational resources, and statistical methods, the processes to produce these finished BS datasets are still computationally expensive (Cheriyadat *et al.*, 2007; Esch *et al.*, 2018a, 2018b), with a single 185km x 180km Landsat 8 scene at the equator containing approximately $2.96 \cdot 10^{10}$ 30m pixels across 8 multispectral bands (excluding the thermal and panchromatic bands).

The lack of time specific and/or comparable built-environment data has not allowed larger questions to be addressed, such as what are the relationships between population distributions and the built environment and the temporal and spatial changes of population in relation to changes in built-environment distribution and built-environment morphology across large extents. Further, more direct applications require time-specific and consistently defined built-environment data, e.g. to define urban or rural (Henderson *et al.*, 2003; Gaughan *et al.*, 2016), and equally time-specific population distributions for planning purposes and to monitor progress of interventions or policy effects (McGranahan, Balk and Anderson, 2007; Patel *et al.*, 2015; Bharti *et al.*, 2016; Linard *et al.*, 2017; Juran *et al.*, 2018; Tatem, 2018). These needs for time specific applications and research, combined with the current temporal limits of BS datasets, have prompted some to interpolate BS extents in a globally consistent manner, producing annually estimated BS extents and expanding the temporal coverage of these BS datasets while maintaining its dataset specific definition of BS (Nieves *et al.*, 2020a).

A larger question accompanying any application, even if modelled extents were found to be accurate within their validation framework, is how these modelled built-environment extents contribute to subsequent modelling applications. To the best of our knowledge, no large-scale assessment of the potential contribution, informative or not, of an urban growth model has been undertaken. This is particularly so for assessing the potential impact of utilising modelled built-environment extents in time-specific modelling population distributions. Lacking time specific built-environment extent data, top-down, i.e. disaggregative, population modelling applications typically utilise the last observed RS-derived built-environment extents (Balk *et al.*, 2004, 2006).

This prompts us to ask whether modelled built-environment extents are more informative than the last observed built-environment extents within a population modelling context. More generally, we want to know if time-specific modelled built-environment extents contribute meaningful information to population models and if this varies across region and time. Additionally, we also want to see the relative contribution of time-specific modelled extents to time-specific RS-derived extents and see if this varies by region and across time.

2. Materials and Methods

To begin to examine how modelled BS could contribute meaningfully to population modelling applications, we examine 2,236 year-and country-specific model objects of the WorldPop "Global Project" (WorldPop - School of Geography and Environmental Science - University of Southampton; *et al.*, 2018), which were used in disaggregative modelling census-based population counts and estimates from 2000-2020. Specifically, we look at a scenario where BS extents were annually interpolated (Nieves *et al.*, 2020a) globally between 2000-2012 and subsequently used as a covariate to predict corresponding annual gridded population surfaces. These models included time-specific modelled BS extents covariate, time-specific RS-derived BS extents covariate, and a BS extents covariate corresponding to the year 2000, allowing us to address our research questions posed. We perform a meta-analysis (Nieves *et al.*, 2017) of the covariate importance of the time-specific modelled BS extents covariate, relative to all other covariates, in modelling population density through a top-down disaggregative framework.

2.1 Study Area

Here we examine 222 countries across the years 2000-2012. Countries were excluded from analysis because they either did not have the BSGM model run (due to resource limitations) or they were modelled using a regional model parameterization, similar to Gaughan et al. (Gaughan *et al.*, 2014), resulting in 172 countries for analysis across 13 years. Regional parametrization precludes any analysis of the country specific importance of any covariates due to the merging of random forest model objects (Table 1) (Nieves *et al.*, 2017). Of specific note was the exclusion of the USA. We excluded it from this analysis because the BS model was not run on its 10 million plus subnational units and large spatial extent due to project resource limitations. For analyses we adopted a regional grouping of countries initially based upon the World Bank's regional groupings (The World Bank, 2020), but modified in some areas based upon economic, historical, developmental, and urbanisation context similarity/dissimilarity (Figure 1). Because the “North American” region only included two modelled countries (Canada and Greenland), we excluded it from further analyses. A full list of countries that were modelled and their region grouping is in Appendix A, Table 1 and a list of countries excluded from our analysis, and the corresponding reason, are in Appendix A, Table 2.

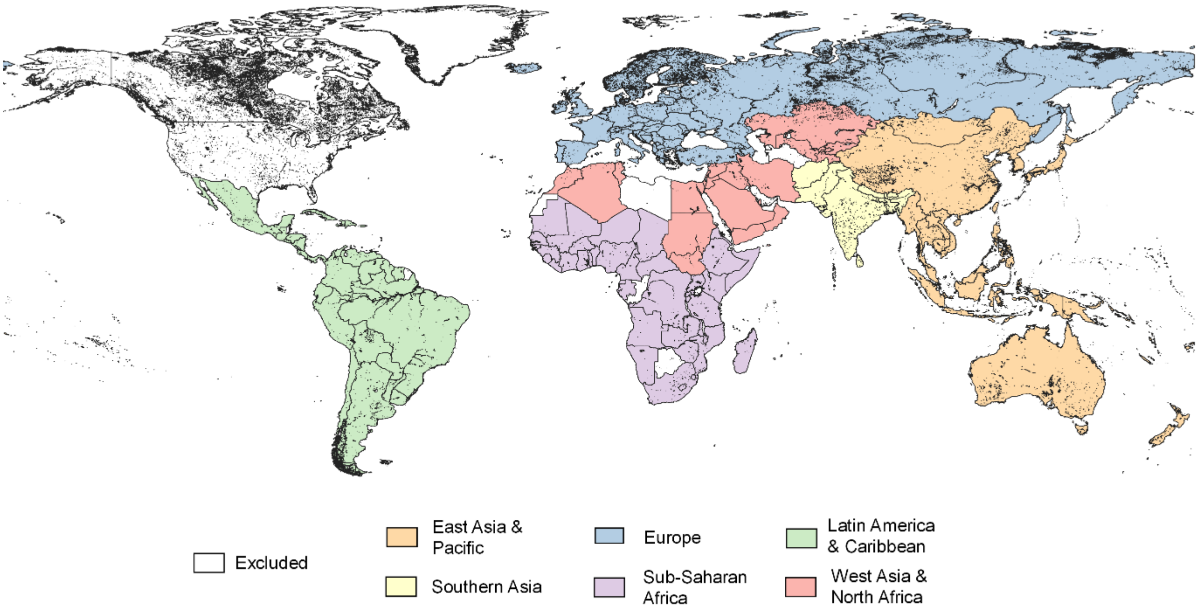


Figure 1. Map of countries included in the meta-analysis and the regional groups used in analyses. See Appendix A, Table 2 for a list of countries excluded from analyses and corresponding exclusion criteria.

2.2 Population Data

Annual estimates of subnational population across the globe were provided by the Center for International Earth Science Information Network (CIESIN) and are based upon the work of Gridded Population of the World, version 4 (GPW, v4). Population counts are based upon censuses and/or official estimates which were interpolated to estimate annual counts, following Doxsey-Whitfield et al. (Doxsey-Whitfield *et al.*, 2015). The subnational unit areas (hereafter simply “unit”) were spatially harmonized and assigned a unique identifier corresponding to a globally consistent grid of harmonized coastlines and international borders, as described in Lloyd et al. (2019).

2.3 Built-Settlement (BS) Data

Built-settlement (BS) (Nieves *et al.*, 2020a) is based upon the definition put forth by Pesaresi et al. (2013, p. 2108), “...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." This was further generalised by Nieves et al. (2020a) to include any datasets attempting to better capture buildings and structures within the above definition while attempting to exclude general impervious surface land cover which lacks a vertical dimension (e.g. roads, runways, parking lots), whether this is achieved through a feature extraction process or from post-processing.

Here we selected a combination of the Global Human Settlement Layer (GHSL) 38m settlement extents for the year 2000 (Pesaresi *et al.*, 2013; Corbane *et al.*, 2017), the "Urban areas" thematic class, class 190, from the ESA CCI land cover 300m global time series for the year 2000 (hereafter ESA) (ESA CCI, 2017), and the Global Urban Footprint (GUF) 72m settlement extents representing circa 2012 (Esch *et al.*, 2013). These data were resampled to 100m and spatially harmonized as detailed in Lloyd et al. (2019), with the ESA data used, in conjunction with the information supplied by the GUF 2012 information, to systematically back-fill missing portions within large settled areas due to imagery availability and atmospheric conditions. The resulting BS extents, for 2000 and 2012, were used as is to derive covariates for use in predicting the annual interpolated BS extents, 2001 through 2011, and for predicting gridded population surfaces, for their corresponding year of representation.

2.4 Geospatial Covariates

We utilised a suite of geospatial covariates in interpolating the annual BS extents as well as disaggregating the annual unit-area population counts into annual gridded population surfaces. All covariates were produced as described in Lloyd et al. (2019), with categorical covariates converted to a continuous covariate, by calculating the Distance-To-nearest-Edge (DTE), for areal type covariates and distance-to-nearest feature calculated for linear and point type covariates. A list of covariates, their original resolution, their source, and a description of them are given in Tables 2 and 3.

Table 2. Table of geospatial covariates used in the modelling of annual BS using the interpolative Built-Settlement Growth Model (BSGMi) per Nieves et al. (2020). Here, representation of BS here is a combination of ESA, GHSL, and GUF as described in Lloyd et al. (2019).

Covariate	Description	Use [a, c]	Time Point(s)	Original Spatial Resolution(s) (arc seconds)	Data Source(s)
Built-settlement [b]	Binary BS extents	Demand Quantification Spatial Allocation	2000 2012	10, 2, & 1	(Pesaresi *et al.*, 2013; Esch *et al.*, 2013; ESA CCI, 2017)
DTE Built-settlement	Distance to the nearest BS edge	Spatial Allocation [c]	2000	10, 2, & 1	(Pesaresi *et al.*, 2013; Esch *et al.*, 2013; ESA CCI, 2017)
Proportion Built-settlement 1,5,10,15	Proportion of pixels that are BS within 1,5,10, or 15 pixel radius	Spatial Allocation [c]	2000	10, 2, & 1	(Pesaresi *et al.*, 2013; Esch *et al.*, 2013; ESA CCI, 2017)
Elevation	Elevation of terrain	Spatial Allocation [c]	2000 [e]	3	(Lehner, Verdin and Jarvis, 2008)
Slope	Slope of terrain	Spatial Allocation [c]	2000 [e]	3	(Lehner, Verdin and Jarvis, 2008)
DTE Protected Areas Category 1	Distance To the nearest Edge (DTE) of level 1 protected area	Spatial Allocation [c]	2012	Vector	(U.N. Enviroment Programme World Conservation Monitoring Centre and IUCN World Commission on Protected Areas, 2015)
Water	Areas of water	Restrictive Mask	2015 [e]	5	(Lamarche *et al.*, 2017)
Subnational Population	Annual population by sub-national units	Demand Quantification	2000 - 2020	Vector	(Doxsey-Whitfield *et al.*, 2015)
Weighted Lights-at-Night (LAN) [d]	Annual lagged and sub-national unit normalised LAN	Spatial Allocation	2000-2011	30	DMSP (Zhang, Pandey and Seto, 2016; Lloyd *et al.*, 2019)

a 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
(see c) 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.

b The binary BS data utilised 2000 and 2012 as observed points in the dasymetric modelling process, but only derived covariates for 2000 were utilised in the random forest as predictive covariates

c Used as predictive covariates in the random forest calculated probabilities of transition

d Readers are referred to Nieves et al. [5] for details on the lagging, normalizing and weighting procedure.

e. Assumed time-invariant

Table 3. Table of geospatial covariates used in the disaggregative modelling of gridded population surfaces.

Covariate	Variable Name(s) in Random Forest	Description	Time Point(s)	Original Spatial Resolution(s) (arc seconds)	Data Source(s)

DTE Built-settlement [a, b]	ghsl_esa_dst;
	bsgm_wpgp_dst
	ghsl_guf_dst;
	ghsl_esa_dst_2000
		Distance To the nearest Edge (DTE) of BS	2000;
			2001-2011;
			2012;
			2001-2012
				10, 2, &1	(Pesaresi *et al.*, 2013; Esch *et al.*, 2013; ESA CCI, 2017; Lloyd *et al.*, 2019)

Elevation	Topo	Elevation of terrain	2000 [e]	3	(Lehner, Verdin and Jarvis, 2008; Lloyd *et al.*, 2019)
Slope	Slope	Slope of terrain	2000 [e]	3	(Lehner, Verdin and Jarvis, 2008; Lloyd *et al.*, 2019)
Lights At Night (LAN)	dmsp; viirs	Annual average of LAN atmospheric radiance	2000-2011; 2012	30	(Earth Observation Group, 2013; Lloyd *et al.*, 2019)
DTE Protected Areas Category 1	wdpa_cat1_dst	Distance To the nearest Edge (DTE) of level 1 protected area	2000-2012	Vector	(U.N. Enviroment Programme World Conservation Monitoring Centre and IUCN World Commission on Protected Areas, 2015; Lloyd *et al.*, 2019)
Water	cciwat_dst	Areas of water to mask areas of model prediction and, for inland bodies of water, as a DTE covariate	---	5	(Lamarche *et al.*, 2017; Lloyd *et al.*, 2019)
Subnational Population	---	Annual population by sub-national units	2000 -2020	Vector	(Doxsey-Whitfield *et al.*, 2015)
ESA CCI Land Cover (LC) Class [c]	ccilc_dst<*class number*>_<*year*>	Distance To nearest Edge (DTE) of individual land cover classes	2000	10	(ESA CCI, 2017; Lloyd *et al.*, 2019)
Distance to OSM [d] Rivers	osmriv_dst	Distance to nearest OSM river feature	2017	Vector	(OpenStreetMap Contributers, 2017; Lloyd *et al.*, 2019)
Distance to OSM [d] Road Intersections	osmint_dst	Distance to nearest OSM road intersection feature	2017	Vector	(OpenStreetMap Contributers, 2017; Lloyd *et al.*, 2019)
Distance to OSM [d] Roads	osmroa_dst	Distance to nearest OSM road feature	2017	Vector	(OpenStreetMap Contributers, 2017; Lloyd *et al.*, 2019)
Average Precipitation	wclin_prec	Mean Precipitation	1950 - 2000	30	(Hijmans *et al.*, 2005; Lloyd *et al.*, 2019)
Average Temperature	wclim_temp	Mean temperature	1950 - 2000	30	(Hijmans *et al.*, 2005; Pezzulo *et al.*, 2017)

a ghsl_esa_dst was only used in the year 2000 population model; bsgm_wpgp_dst was derived from the BSGM predicted extents and used for years 2001-2011; ghsl_guf_dst was used for the year 2012

b ghsl_esa_dst_2000 is identical to ghsl_esa_dst, but was included as a covariate in all models from 2001 onward to avoid unrealistic population distributions as seen in multitemporal modelling within Gaughan et al. (2016)

c Some classes were collapsed: 10-30 → 11; 40-120 → 40; 150-153 → 150; 160-180 → 160 (Sorichetta *et al.*, 2015)

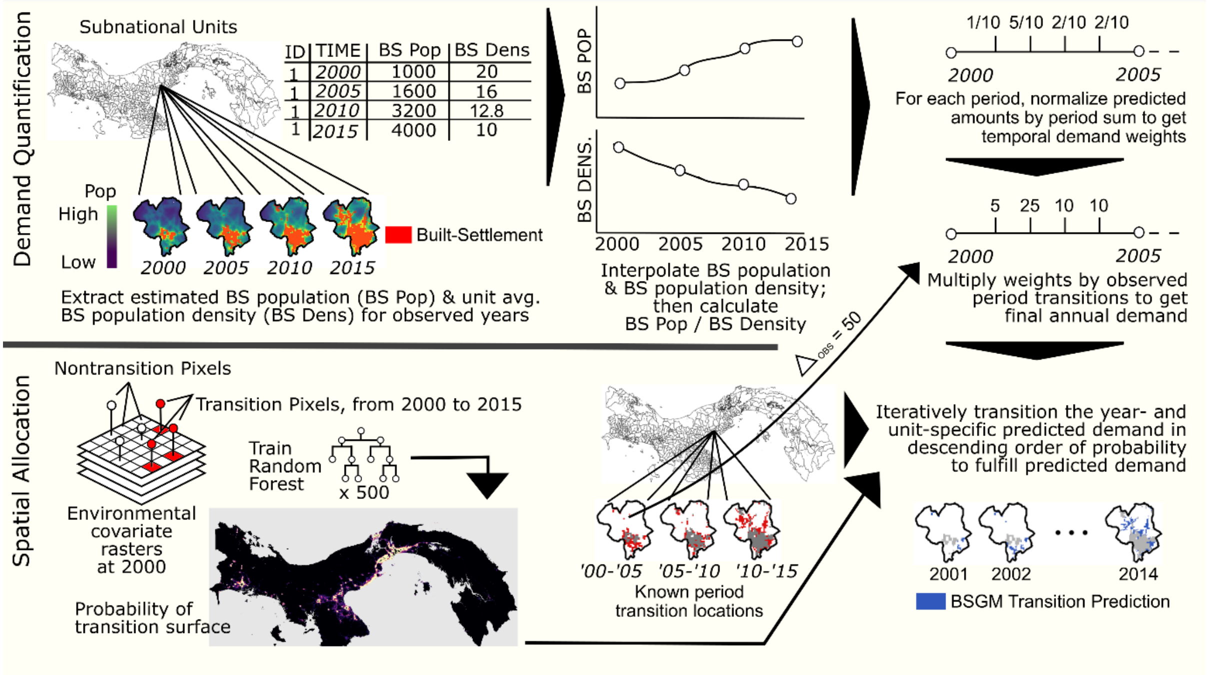
d OpenStreetMap (OSM)

e Assumed time-invariant

2.5 Methods

2.5.1 Built-Settlement Growth Model interpolation (BSGMi)

The BSGMi is a top-down modelling framework which disaggregates observed numbers of non-BS-to-BS transitions from coarser spatial and temporal resolutions to finer spatio-temporal resolutions using ancillary data (Nieves *et al.*, 2020a). It can be generally thought of as having two primary components: a Demand Quantification component and a Spatial Allocation component (Figure 2) (Nieves *et al.*, 2020a).



ID	TIME	BS Pop	BS Dens
1	2000	1000	20
1	2005	1600	16
1	2010	3200	12.8
1	2015	4000	10

Figure 2. Generalised BSGMi process diagram from Nieves et al. (2020).

Assume we are given a time period of with at least two observations of BS extents, typically derived from remote sensing imagery, and corresponding estimated time- and unit-specific population found spatially coincident with the BS extents (Nieves *et al.*, 2020a). At regularly spaced intervals between the two or more observations, we can interpolate the BS population using unit-specific logistic growth curves estimate unit-level BS population (Figure 2) (Nieves *et al.*, 2020a). Similarly, we can use natural cubic splines to interpolate unit-level changes in BS population density (Figure 2) (Nieves *et al.*, 2020a). We then use the relative unit-level changes in interpolated BS population and BS population density to derive time- and unit-specific weights (representing unit-level non-BS-to-BS transition demand) which we use to temporally disaggregate the observed non-BS-to-BS transitions from the larger time period to the finer regularly spaced intervals, in this case years, between the two or more observations (Figure 2)

(Nieves *et al.*, 2020a). This has the benefit of preserving agreement with the observed points (Mennis, 2003; Mennis and Hultgren, 2006; Nieves *et al.*, 2020a).

Once the number of transitions at the desired temporal level have been estimated, we move to the Spatial Allocation component of the modelling framework (Figure 2) (Nieves *et al.*, 2020a). Here we utilised a Random Forest (RF) model (Breiman, 2001a; Liaw and Wiener, 2002), using predictive covariates listed in Table 2, to predict the pixel level probability of a non-BS-to-BS transition occurring between any two observed extent points (Nieves *et al.*, 2020a). This represents the period level probability of transitioning and is further modified by using annual differences in lights-at-night (LAN) radiance values that are rescaled based upon the value distribution within their respective subnational units (Nieves *et al.*, 2020a). The values are rescaled in such a way that pixels with greater unit-relative increases in LAN brightness are thought to indicate a higher probability of transitioning and vice versa (Nieves *et al.*, 2020a). Multiplying the RF pixel probabilities by the corresponding LAN weights produces year-specific probability surfaces which are then used, on a unit by unit basis, to iteratively disaggregate the year-specific predicted transitions (from the Demand Quantification component) across space (Figure 2) (Nieves *et al.*, 2020a). This produces a gridded time series of BS spatial extents between every observation given (Figure 2).

Validation of the BSGMi framework across four sample countries at 100m pixel resolution, given 4 observed years and predicting for twelve years, showed consistent performance across a variety of environments and contexts with the majority of interpolated years having a pixel level accuracy of greater than 80 percent (range 57 to 99 percent) (Nieves *et al.*, 2020a). However, the BSGMi framework utilised by the Global Project was an early version and differed from the version validated by Nieves et al. (2020a) in two systemic ways: both the BS population and BS population densities were interpolated using unit-specific exponential growth/decay curves and the model was fit using only information from two time points at a time. This would likely result in an increased likelihood of overfitting for the BS population density across time, i.e. interpolated using information from two points rather than more than two, and a shifting of transitions to later in the time period due to the exponential shape. Nieves et al. (2020a) found the model tended to predict transitions late so the latter, speculated, effect of having exponential assumption might mitigate this, but the magnitude and effect are unclear without further work. Additionally, given that the BSGMi is an interpolative method, it is highly sensitive to the selected representation of BS selected as input. Nieves et al. (2020a) utilised the, originally, coarser 300m ESA CCI "urban" land cover dataset given its annual coverage allowed for holdout samples for validation whereas, here, we are using a combination of the relatively, originally, finer resolution 38m GHSL and 72m GUF data products that have been backfilled by the ESA CCI land cover data per Lloyd et al. (2019).

Despite these differences, the binary representation of the annual BS extents produced using the BSGMi were then converted into a continuous representation of the Distance-To-nearest-Edge (DTE) of BS. This conversion to continuous distances and the fact the population models examined in this study are at the subnational unit-level, requiring us to take the unit-average DTE of BS, this does effectively smooth any of the more frequent and smaller differences that would likely result, at various scales, due to the aforementioned differences between the validated BSGMi framework (Nieves *et al.*, 2020a) and the early BSGMi framework we utilise here.

2.5.2 Top-down RF Population Disaggregation

The Global Project utilised a top-down RF informed dasymetric population disaggregation to distribute unit-level census-based population counts to pixel level (100m) population count estimates (Gaughan *et al.*, 2014, 2016; Sorichetta *et al.*, 2015; Stevens *et al.*, 2015). RFs were chosen due to their automatability, scalability, ability to capture complex interactions and non-linear phenomena, and robustness to small samples and noise (Farror and Glauber, 1967; Breiman, 2001a; Rodriguez-Galiano *et al.*, 2012). This modelling approach was applied on a country-by-country basis using a suite of globally harmonized and time-specific, or assumed temporally invariant, geospatial covariates which were aggregated by calculating the average of values within each subnational unit prior to being input to the RF (Figure 3) (Gaughan *et al.*, 2014, 2016; Sorichetta *et al.*, 2015; Stevens *et al.*, 2015).

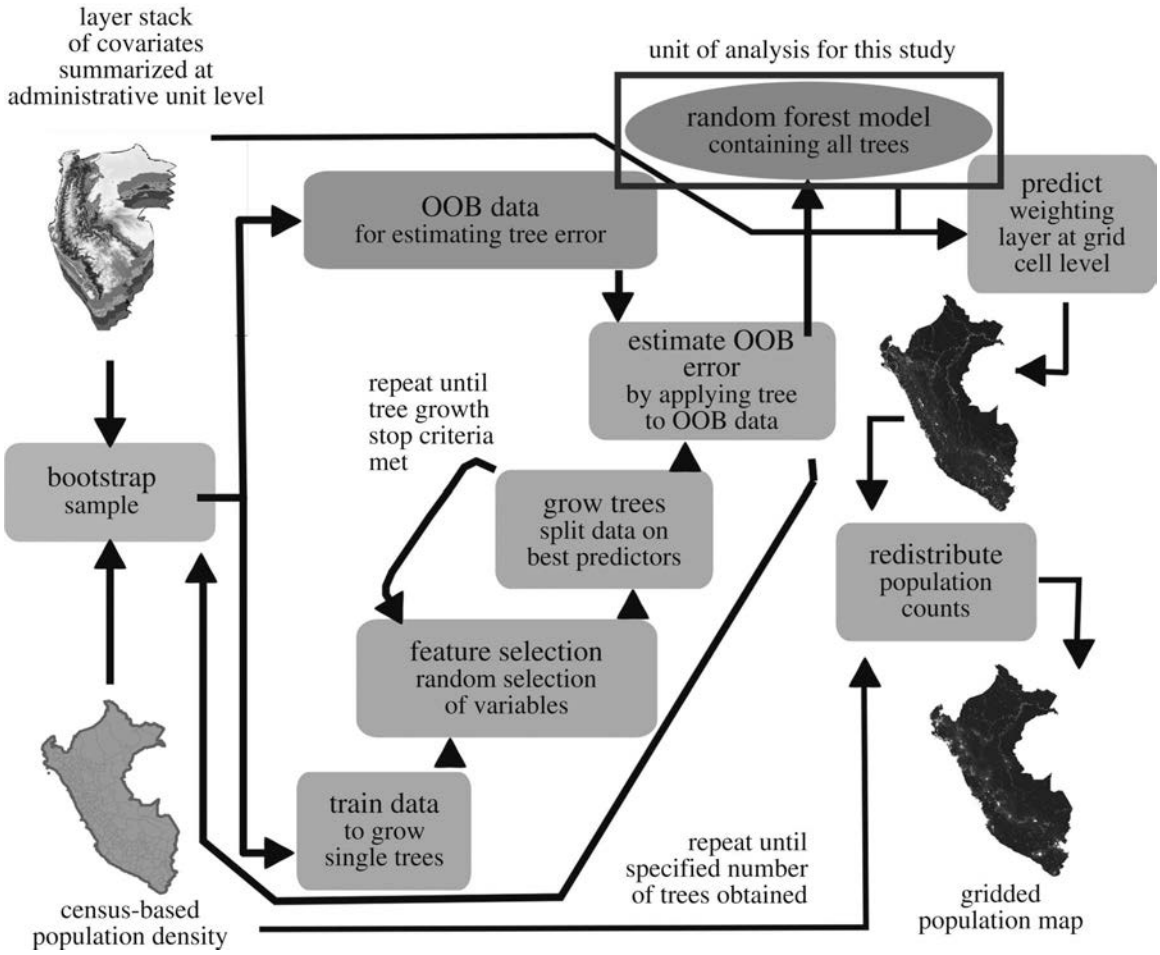

layer stack of covariates summarized at administrative unit level
unit of analysis for this study
random forest model containing all trees
OOB data for estimating tree error
predict weighting layer at grid cell level
estimate OOB error by applying tree to OOB data
repeat until tree growth stop criteria met
bootstrap sample
grow trees split data on best predictors
redistribute population counts
feature selection random selection of variables
train data to grow single trees
repeat until specified number of trees obtained
census-based population density
gridded population map

Figure 3. Generalised diagram of the RF-informed dasymetric disaggregation of population counts from subnational units to a given pixel level. Figure from Nieves et al. (2017).

While trained at the unit-level, the RF is then used to predict population density at the pixel level (100m); we use these predictions as unit-relative weights to disaggregate the corresponding unit population count to pixel-level population counts while ensuring that the sum of pixel-level values sums up to the original unit-level count (Figure 3) (Gaughan *et al.*, 2014, 2016; Sorichetta *et al.*, 2015; Stevens *et al.*, 2015). Each year's population disaggregation was done independently of the others.

RF models are a class of ensemble model where many "weak" classification and regression trees are combined through voting or averaging to produce more robust predictions (Breiman, 2001a). In this study, we utilised the *tunerf* function (Liaw and Wiener, 2002) to determine the optimal number of covariates to examine at each iterative split and carry out an iterative covariate selection process, per Stevens et al. (2015), to remove any covariates with an average Percent Increase in the Mean Squared Error (Per.Inc.MSE) less than or equal to zero (Stevens *et al.*, 2015). The Per.Inc.MSE is an internal cross validation metric of covariate importance that is calculated by permutating the covariate information, preserving all other covariate information, and averaging the percent increase in the mean squared error across all trees in the RF when withheld “Out of Bag” (OOB) (Breiman, 1996, 2001a) data is compared to the RF predictions. For further details on constructing RF models, bagging, and covariate selection and splitting in a random forest we refer readers to (Breiman, 1996, 2001a; Liaw and Wiener, 2002; Strobl *et al.*, 2007, 2008).

However, Per.Inc.MSE is a relative, model specific, measure of importance that is highly conditional upon the other present covariates (Breiman, 2001a), presenting a challenge for using this metric when attempting to compare, even with a static set of covariates, the covariate importances across models (Nieves *et al.*, 2017). Additionally, while it is generally understood that predictions produced by a RF are resilient to the issue of multi-collinearity, it does not preclude multi-collinearity from affecting the relative covariate importances within a given model. For instance, as is the case with the models examined here, if you have multiple representations of BS covariates in the model, with each covariate having partially overlapping fields of capture in the information space (i.e. multi-collinearity), and all are retained in the model, then the magnitude of the Per.Inc.MSE of will be “stolen” from the most important covariate (Breiman, 2001a). However, the relative ranking of the correlated covariates will be proportional to their frequency of utilization as splitting criteria across all trees, i.e. the most important covariate of the correlated covariates will still have the highest Per.Inc.MSE, it will just be of a smaller magnitude than without the inclusion of the correlated covariates in the RF.

2.6 Analyses

Given the potential difficulties of comparing covariate importance across independent RF models, we adopt the Weighted Importance Rank (WIR) from Nieves et al. (2017) to facilitate our comparison of covariate importance across country- and time-specific RF population models. The WIR accounts for the potentially different number of covariates in each model, resulting from the covariate selection, by taking the ranking covariates within a given model by descending importance and dividing this rank by the total number of covariates in the model (Equation 1) (Nieves *et al.*, 2017).

$$WIR = \frac{within-model\ ranked\ importance}{total\ number\ of\ covariates\ in\ model} \qquad [1]$$

This results in a value between 0 and 1, with the most important covariate having a value of 0 and the least important having a value of 1(Nieves *et al.*, 2017). Hereafter, when referring to covariate importance, we are referring to the WIR as opposed to Per.Inc.MSE.

We collected all the RF model objects (*n* = 2236) produced in the modelling of population for the years 2000-2012, extracted the covariate importances (Per.Inc.MSE) into a data table, transformed the importances to WIR values, and assigned each country a label corresponding to their region (Figure 1). Similar to Nieves et al. (2017), we discovered the non-normal distributions of covariate importance data and, accordingly, adopted non-parametric statistical methods in conjunction with visual analyses. Using Kruskall-Wallis tests (Kruskal and Wallis, 1952; Rosner, 2011), we tested for significant differences in the variable importance distributions of the BSGMi derived covariate: (i) between years 2001-2011 across all countries and, (ii), between countries grouped by regions (Figure 1), across all years 2001-2011. Additionally, to determine if the year-specific BSGMi-derived covariate was adding additional information to the models for years 2001-2011, we calculated the differences in WIR distributions: (i) between the BSGMi-derived covariate and the BS extents at the year 2000 (GHSL-ESA 2000), (ii) between the BSGMi-derived covariate and the annual RS-derived "urban areas" extents (ESA Annual), and, (iii) between the GHSL-ESA 2000 covariate ad the ESA Annual covariate. We then carried out one-sample Wilcoxon rank sum tests (Wilcoxon, 1945) to determine if there was a significant difference in the distributions of the WIR difference and a zero-median difference.

All Kruskall-Wallis and Wilcoxon rank sum tests were carried out with $\alpha = 0.05$ and, if significant results were found for the Kruskall-Wallis tests, these were followed up with *post hoc* Dunn tests with Holm correction for multiple outcomes (Dunn, 1964; Holm, 1979). Wilcoxon rank sum tests were adjusted for multiple outcomes as well using Holm's correction. All models were carried out using the R statistical environment 3.4.2 (R Core Team, 2017) and analyses were produced using the R statistical environment 3.6.0 (R Core Team, 2019). All code, tabular data, and full test results are included in the supplementary materials.

3. Results

Globally, across all years in the study period, we can see very consistent patterns of covariate importance. For clarity, we focus on five years (2000, 2003, 2006, 2009, 2012) and the four most important covariates (Lights-At-Night covariates, the BSGMi-derived covariate, the ESA Annual covariate, and the GHSL ESA 2000 covariate), hereafter. Based on the median WIR value, the lights-at-night (LAN) covariate is the most important covariate across all years (Figure 4). For 2001 through 2011, the second, third, and fourth most important covariates are, respectively, the BSGMi-derived covariate (BSGMi), the annual RS-derived ESA covariate (ESA Annual), and the RS-derived covariate representing 2000 BS extents (GHSL-ESA 2000) (Figure 4). For the BSGMi covariate, we show that the variance decreases, and the median importance increases (smaller WIR value) with time, converging towards the 2012 GHSL GUF covariate's distribution, which is what we would expect if the BSGMi model is interpolating accurately. Further, the distribution of the WIRs of the BSGMi-derived extents covariate appear to show consistency from one year to the next with an overall trend of decreasing WIR variance as the year becomes closer to 2012. At the global level, between years, there is no significant difference in the WIR distributions of the BSGM derived covariate ($X^2 = 15.1$, df = 10, $p = 0.13$; full results in supplementary materials).

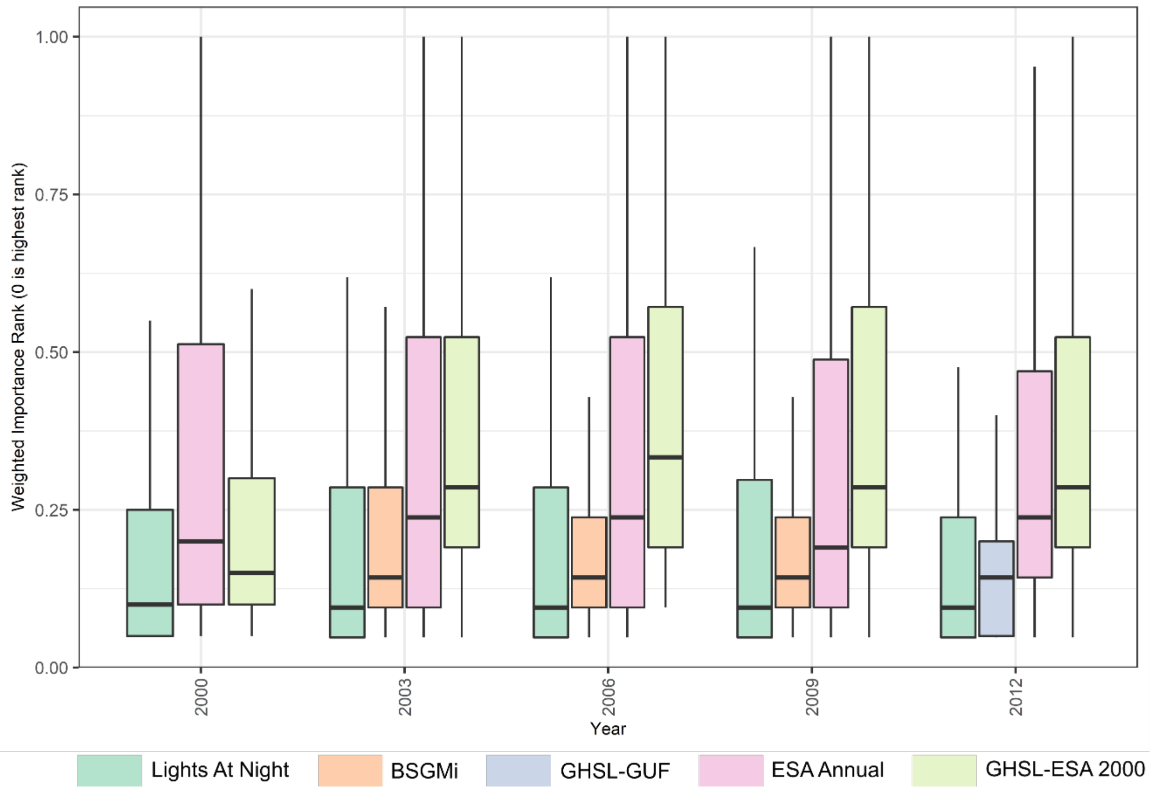


Figure 4. Boxplots of the weighted importance rank (WIR) of the four most important covariates in each year's random forest model. WIR value distributions are shown for all countries by year with the median shown as a black line dividing the interquartile range (IQR, shown as the boxes) and 1.5 * the IQR being represented by the "whiskers" of the plots.

Looking only at the distributions of the BS-related covariates, we plotted the WIR boxplots by year and region in Figure 5. Within a given region, it would appear that there is generally consistent performance of the BSGMi-derived covariate with some regions exhibiting a slight temporal trend between 2000 and 2012, showing the large differences in GHSL dominated information (2000) and GUF dominated (2012) information provided to the RF (Figure 5). A commonality, within most regions, would appear to be that the highest variance in WIR is seen near the midpoint of the interpolation period (2006) where we would expect performance of the BSGMi to be the worst or most variable (Figure 5).

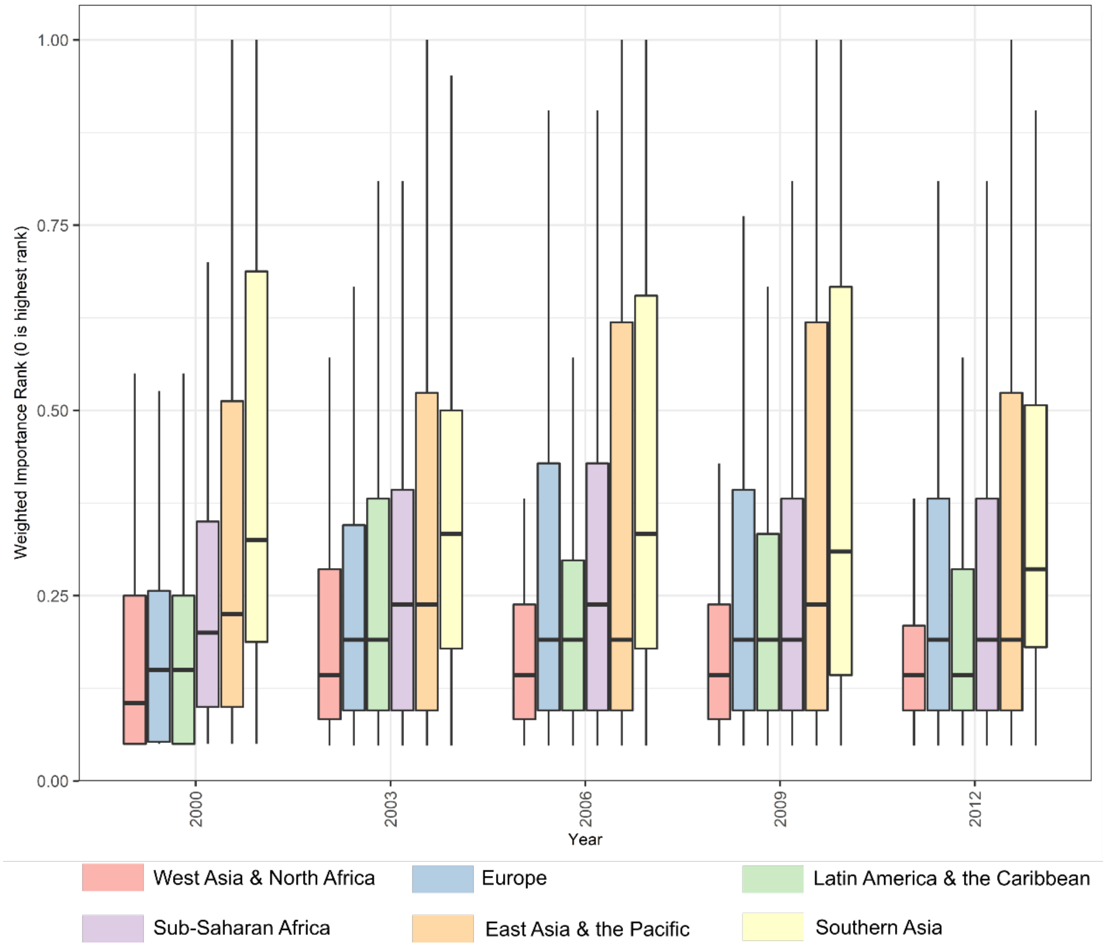


Figure 5. Boxplots of the weighted importance rank (WIR) of BS-related covariates in each countries' random forest model, grouped by region and plotted by year with the median shown as a black line dividing the interquartile range (IQR, shown as the boxes) and 1.5 * the IQR being represented by the "whiskers" of the plots. The BS-related covariate represented in 2000, 2003-2009, and 2012 are, respectively, the GHSL-ESA 2000 covariate, the BSGMi-derived covariate, and the GUF-GHSL covariate.

We plotted the WIR difference between all pairwise combinations of the three covariates of interest and tested their distributions, across all years for each region, to determine if they were significantly different from a distribution with a median WIR difference of 0, i.e. the covariates contribute the same amount of importance (Figure 6, Table 5). When testing for significance, data were aggregated across years 2001-2011 and grouped by region. We show that across all regions the year-specific BSGMi covariate was contributing significantly more importance (p < 0.00 for all regions) to the RF model than the last observed GHSL-ESA 2000 covariate. The largest difference for this is seen in the "South Asia" and "East Asia & the Pacific" regions. When compared to the ESA Annual covariate, the BSGMi covariate is contributing significantly more importance to the RF model in all regions (*p* < 0.00) except “Europe” (*p* = 0.99). Examining the differences between the GHSL-ESA 2000 and the ESA Annual WIR values, we see that the ESA Annual data is contributing significantly more importance in all regions (*p < 0.00*) except the “East Asia & the Pacific” (*p* = 0.14) and the “West Asia & North Africa” regions (*p* = 0.77).

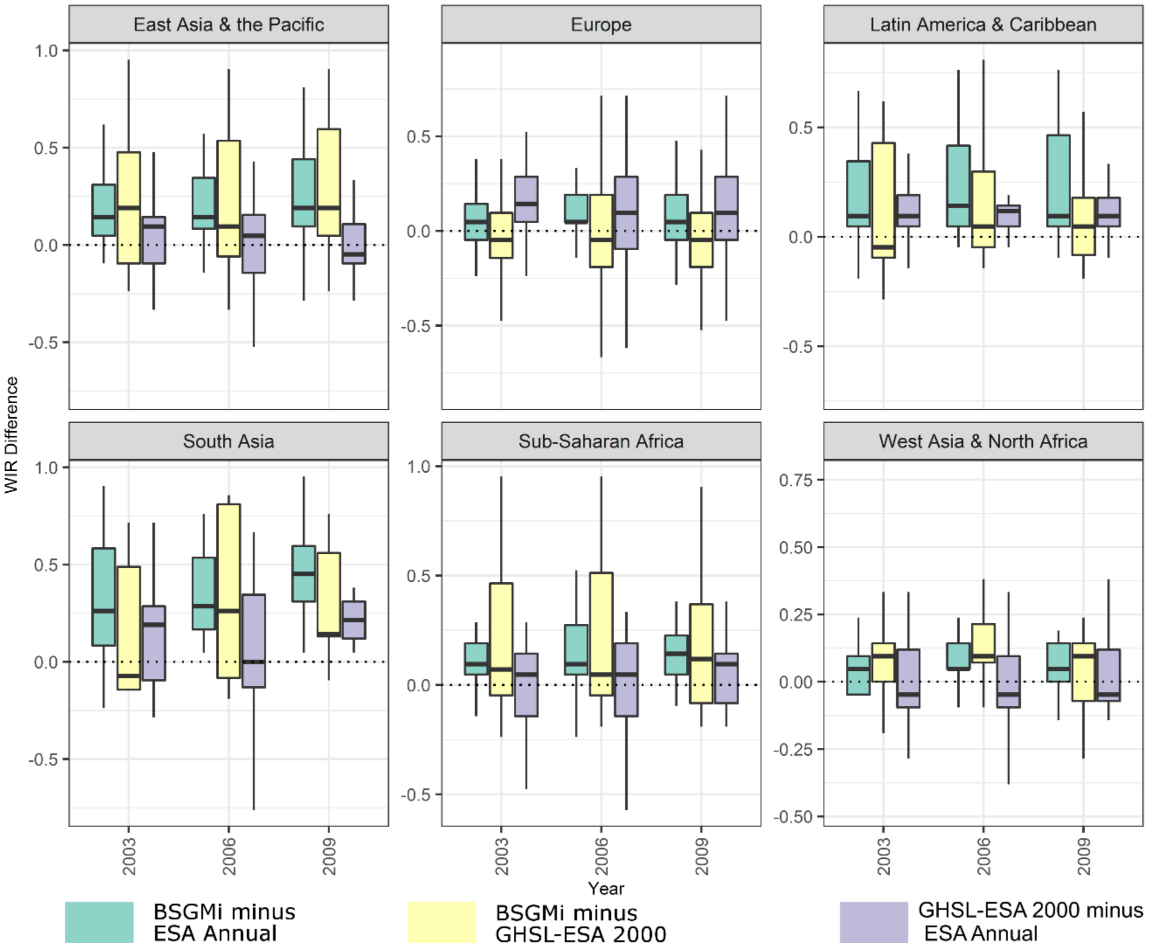


Figure 6. Box plot of WIR difference between the GHSL-ESA 2000 covariate and the year-specific BSGMi-derived covariate, the ESA Annual covariate and the year-specific BSGMi-derived covariate, and the GHSL-ESA 2000 covariate and the ESA Annual covariate. For each comparison, positive WIR differences indicate that the former of the pair was less important than the latter and negative values indicate the opposite. Results for all years are included in the supplementary materials.

Table 5. Adjusted p-values of Wilcoxon one sample test with Holm correction for examining significant differences in covariate importance as measured by the Weighted Importance Rank (WIR). Data was aggregated across years 2001-2011 and grouped by region. Null hypothesis being that the median WIR difference of a given comparison was equal to zero. Significant differences are shaded for emphasis. Full results are provided in the supplementary materials.

WIR Differences	East Asia & the Pacific	Europe	Latin America& the Caribbean	Southern Asia	Sub-Saharan Africa	West Asia & North Africa
GHSL ESA 2000 minus BSGMi	< 0.00	< 0.00	< 0.00	< 0.00	< 0.00	< 0.00
ESA Annual minus BSGMi	< 0.00	0.99	< 0.00	< 0.00	< 0.00	< 0.00
GHSL ESA 2000 minus ESA Annual	0.14	< 0.00	< 0.00	< 0.00	< 0.00	0.77

4. Discussion

We have shown that interpolated year-specific BS-extent data, using the BSGMi framework, is a consistently important predictor of population density globally and across time. Specifically, the BSGMi-derived covariate was consistently second most important, behind year-specific lights at night data. Even though both the lights at night data and the BSGMi data are given to the model as continuous covariates. essentially, the BS-derived covariates only indicate presence and absence of BS while lights at night can capture presence, absence, and intensity of BS presence (Small *et al.*, 2011). However, the year-specific RS-based BS representation (ESA Annual) and the previously observed RS-based BS covariate (GHSL-ESA 2000) are still important (Figures 4 & 5) and can give relative indications of how the chosen BS representation and the BSGMi perform within regions. However, for any given region, these differences in importance were stable across time (Figure 6). Overall, BSGMi interpolated extents increase the information in these population models and, with the other RS-derived covariates, likely better capture the BS-information space as related to population density than any one covariate does alone.

Regardless of the magnitude of the importance or relative importance, a key point is that the BSGMi-derived covariate was always retained in models that it was introduced to and consistently contributed significantly more importance to the models than the other BS representations, across most regions. The fact that all of the representations of BS were consistently the 2nd through 4th most important covariates across all years supports previous importance findings (Nieves *et al.*, 2017) and reemphasizes that utilising multiple representations of

BS results in more accurate disaggregative population modelling (Reed *et al.*, 2018).

We would expect a year-specific BS covariate to contribute significantly more information than a previously observed BS covariate, which was largely supported by the findings in Figure 6 and Table 5. The exceptions in “East Asia & the Pacific” and “West Asia & North Africa” could be explained by several factors: (i) large and or few subnational units, (ii) lack of suitable, e.g. cloud free imagery for these optically based datasets, and/or, (iii) greater difficulty in urban feature extraction within arid regions (i.e. similar radiometric signature between buildings and bare soil) contributing to greater noise in the population density-BS relationship fit by the RF. This could potentially explain the relatively poorer importance contribution of the BSGMi covariate in the “East Asia and the Pacific” and the “South Asia” regions (Figure 5). Additionally, it is important to note that this study uses the original GHSL as a part of its input BS representation and, therefore, it is currently unclear if the newer versions (Corbane *et al.*, 2017), which leverage the increased resolution and different radiometric capture of the Sentinel platforms, would change these findings (Figure 6 and Table 5). The other notable result of Figure 6 and Table 5, the lack of significant difference between the ESA Annual covariate and the BSGMi covariate, could be potentially explained by: (i) the ESA data does rather well within Europe’s dense and well-defined BS extents, and, (ii) those BS extents do not change as much as other regions, i.e. the non-BS to-BS transition prevalence is low so the BSGMi model does relatively worse than in a high transition area (Nieves *et al.*, 2020a). Regardless, it is important to note that the results of Figure 6 and Table 5 are relative and that all the covariate representations of BS were found to be highly important to the RF model of population density.

From previous work (Nieves *et al.*, 2020a), there is little doubt that the BSGMi is picking up true BS extents that, in turn, drives this increased importance. However, the regional differences can more generally be attributed to the chosen RS-derived BS extents input into the BSGMi framework, the quality of the input population data, and the size and configuration of the subnational units used in both the BSGMi and the population modelling method used here (Openshaw, 1984; Stevens *et al.*, 2015; Nieves *et al.*, 2017; Nieves *et al.*, 2020a). To investigate if different underlying structures of causal relationships between population and BS exist, and to then quantify them, a different research framework and modelling approach would be necessary, i.e. an explanatory modelling framework as opposed to a predictive one (Breiman, 2001b; Shmueli, 2010), would be necessary.

Nieves et al. (2020a) suggested that end users of the BSGMi modelling framework check the model outputs for end use suitability and accuracy. The regional differences in the WIR of the BSGMi-derived covariates (Figure 5) reinforce that it is important that users of any modelled BS extents examine them for their use-specific and study area-specific suitability as no model framework is likely to excel in all scenarios. These observed WIR differences can be due to pre-existing differences in the suitability of the input BS representation or due to model-induced uncertainty and error, but in an applied context, the origin is of secondary importance to knowing of its existence.

These findings are for these specific representations of BS and the importances are contingent upon the set of covariates provided (Breiman, 2001a). We would hypothesize that if we were to include the BSGMi-derived covariate as the only representation of BS in the RF models, acknowledging that within a RF correlated variables “take” importance away from each other, there is a possibility that it could surpass the LAN covariate for most important. But, this awaits further study. Further, while here we explored the importance of the BSGMi-derived and other BS-based covariates at the subnational unit level, how this translates into the modelled population distributions and their accuracies remains an open question. We would like to think that having more important covariates at the subnational level would result in more accurate pixel-level disaggregations, but the issues of scale and other inputs into the model make any speculation tenuous, at best. Lastly, the Nieves et al. (2020) validation of the BSGMi framework was with an originally coarser representation of BS (300m ESA CCI landcover) and the authors queried whether the assumed relationships of the framework would hold with originally finer scale input BS extents given their findings and previous findings under a different framework (Tayyebi *et al.*, 2013). While this study does not perform a pixel-based validation of the BSGMi, here we have shown that using originally finer scale input BS extents can produce derived data products that were found to be informative for applications, causing us to speculate that that the framework assumptions do hold. However, whether that indicates the pixel-level BSGMi outputs can be utilised without aggregation, as we have done here for our end use, remains unclear.

Within the population models analysed here, the “last observed” time specific RS-derived BS extents that was originally high-resolution (≤100m) was limited to the year 2000. Therefore, our findings related to importance as compared to the “last observed” would likely change, at a minimum, in magnitude were the “last observed” year to be different, dynamic, or to include multiple “last observed” BS extents. While Gaughan et al. (2016) found that including previous BS extents were important in creating temporally comparable population surfaces when performing top-down modelling, there is no current information regarding at what temporal lag the information contributed is maximized and how many previous representations should be included.

5. Conclusions

Here we tested the utility of the modelled BS extents in a population-modelling scenario across 172 countries and 13 years. Globally, we found that modelled BS extents are consistently the second most important predictor of population density, even when the previous RS-derived BS extents and time-specific BS-extents were included in the model. However, regional variation exists in the importance of the modelled BS extents, but its cause is multifactorial and still unclear. Additionally, there were many cases where the time-specific RS-derived covariate, originally having a coarser spatial resolution, was more important than the high-resolution modelled BS extents and/or the high-resolution previously observed RS-derived extents. Combined with the fact that all covariates were retained in the final models, this would suggest that while modelled BS extents are informative, they are best used in conjunction with other representations of BS when modelling population.

These findings are specific to the spatial scale and zonal configuration of the subnational units used. Future work examining the impact of the scale of the subnational units on both the BS modelling and RF-informed dasymetric modelling should be conducted, although some previous work would indicate that smaller units leads to more accurate models (Andrea E. Gaughan *et al.*, 2014). While this study has shown that the BS modelled extents are important at the subnational unit level, future work should examine how the BS modelled extents affect the pixel level predictions and smaller area population predictions in this top-down modelling framework. Additionally, research into the number of previous extents to include in the population modelling as well as the effect of its temporal lag on population predictions should be investigated.

Chapter 6 Conclusions

The Built-Settlement Growth Model (BSGM) framework provides time-specific and spatially explicit predictions of built-settlement (BS) expansion by leveraging information within subnational changes in population and population density and corresponding environmental covariates. This thesis provides the conceptual and programmatic framework for interpolating and extrapolating BS extents by using relative changes in population at the subnational level (Chapter 3 and Chapter 4). It presents evidence of the accuracy of this framework across a diverse set of physical, socio-cultural, historical, and urban morphological environments. The interpolation framework filled in temporal gaps of the global measurement of BS, which then allowed for the quantification of BS and BS population trajectories (Chapter 3). The quantification of these tens of millions of subnational trajectories allowed for data-driven, subnational extrapolation of near-future BS extents (Chapter 4). Further, this thesis shows that these modelled near-future BS extents, in some situations, can outperform the RS-derived BS extent validation dataset (Chapter 4). And, lastly, this thesis demonstrates that these modelled BS extents have applied utility in informing disaggregative population modelling (Chapter 5).

A significant finding of this thesis is that relative changes in subnational population can be sufficient for predicting the timing and magnitude of BS expansion at high spatial resolutions (Chapter 3 and Chapter 4). Previous urban growth models have largely focused upon economic measures in predicting the demand for urban expansion (Chapter 1). Furthermore, the few previous global/continental urban growth models that incorporated population within their demand quantification components either did not assess the accuracy of predictions at the level of prediction, i.e. pixels (Chapter 1). While the role of economic factors should not be ignored, subnational economic data are typically not available in low- to middle-income contexts, much less at multiple time points. Therefore, quantitatively supporting the finding that relative changes in population can produce accurate predictions of BS expansion greatly expands the applicability of my framework (Chapter 3 and Chapter 4). Additionally, this thesis presents evidence that these findings are globally applicable (Chapter 5).

This thesis also demonstrates that these modelled built-settlement extents are useful and informative for population modelling end applications (Chapter 5). The modelled BS covariates were consistently the second most important predictor of population density across the globe, regularly outperforming time-specific RS-derived BS covariates (Chapter 5). This likely is due to population data better predicting the presence and or expansion of settlement in areas where the extraction of built features from imagery is difficult or imagery availability/quality is limited.

In a technical sense, this thesis represents a significant step forward in urban growth modelling. It introduces the first globally applicable BS modelling framework that:

i) Allows for subnational variation to drive the larger scale model
ii) Has very little data requirements, i.e. only binary extents and subnational population data at a minimum
iii) Requires no user assumptions, “expert knowledge”, or other *a priori* parameters
iv) Has undergone explicit pixel level validation of the extent predictions
v) Allows for any binary representation of urban and any subnational population dataset as input.

These technical advances presented in this thesis provide foundations for further advances in urban growth modelling which in turn can drive better understanding of urban and population dynamics, such as per capita built land use (Verburg and Overmars, 2009; Angel *et al.*, 2011; Seto *et al.*, 2011; Balk *et al.*, 2018), population density changes in settlements over time, and trends in population and settlement distribution in response to climate change (McGranahan, Balk and Anderson, 2007). Specifically, the emphasis of this modelling framework on preserving subnational variation and using empirical subnational trajectories provide a data set rich with possibilities for secondary analyses. One such analysis could include better determining, at finer spatial and temporal scales, if urbanised areas are generally experiencing a decrease in population density (Angel *et al.*, 2011). Another could simply investigate what environmental covariates were the most important predictors of areas transitioning from non-BS to BS; potentially giving insight into larger land use dynamics across time.

This thesis puts forth and explicitly validates these novel modelling frameworks (Chapter 3, Chapter 4, and Chapter 5); something that is not always possible or done with past urban modelling work (Goldewijk, Beusen and Janssen, 2010; Angel *et al.*, 2011; Seto *et al.*, 2011; K C Seto, Guneralp and Hutyra, 2012; Linard, Gilbert and Tatem, 2013; Tayyebi *et al.*, 2013) (Chapter 1). Moreover, I exhibited the practicality of scaling such flexible frameworks by modelling BS expansion globally over a 20-year period, producing outputs that demonstrated informative utility in population modelling and other applied contexts (Chapter 5).

The data set outputs from this thesis include annual 100m resolution global BS extents from 2000 through 2020, used to produce an annual 100m resolution global population maps for the same time period. These data have already had substantial use in the humanitarian response and development community. With over 7,900 downloads of the BSGM modelled datasets since March 2019 and over 34,000 downloads of the population datasets that utilised the BSGM modelled outputs as covariates. Outside of these downloads directly from the WorldPop website, the modelled BS extents and derived population datasets are also distributed on the Humanitarian Data Exchange (HDX) where data is often used in research, disaster response (e.g. building damage assessment, internally displaced persons and refugee monitoring), and humanitarian development. These datasets are also utilised by the Institute for Health Metric and Evaluation (IHME) in better estimating the denominators of rates of disease and disease burden, such as the "Local Burden of Disease" project (https://www.healthdata.org/lbd). Most recently, the IHME have used the WorldPop Global Project's 2020 population density projections, which are largely driven by the BSGMe projections, in their COVID-19 model (Institute for Health Metrics and Evaluation, 2020), further emphasising the real need for globally consistent short-term forecasts of population and built-settlement. Additionally, the GRID3 project (https://grid3.org/) uses the BS extents in their efforts for modelling country populations in the absence of a census and the GridSample tool (https://gridsample.org/) uses the BS extents in its effort to better guide representative survey sampling. Academic researchers and governmental institutions in other countries, including Belgium (MAUPP project, https://maupp.ulb.ac.be/) and South Africa (South African National Space Agency), have requested the BSGM framework for further applications and research related to better mapping spatial population changes over time and applications of the models to their own urban and BS datasets.

6.1 Limitations and Caveats

Considering these accomplishments, these works still have limits and caveats that need to be considered. Many of the limits of the frameworks, methods, and analyses are presented in the individual papers, but some key general limits are discussed here.

In this thesis, the BSGM framework uses BS population estimates derived from modelled population, which uses environmental covariates, to estimate changes in the built-environment. This process leverages correlations between population and the built-environment to make predictions about both populations and, separately, about the built-environment. While the two-stage hierarchical nature of the framework, where population is used to estimate demand at one spatial scale and the demand is met at a finer spatial scale with no input from population, limits concerns of endogeneity, the concerns are not entirely absent. In the configurations used here, these modelling frameworks are not suited for making inferences about causality of these changes. Models either explain (i.e. infer causality) or predict (Breiman, 2001b; Shmueli, 2010). Issues of endogeneity or "circular inference" become a cause of concern when a model is meant to be predictive but is used to infer something about the causal relationships between covariate and outcome, e.g. the shape (linear, quadratic, etc.) or magnitude of the causal effect between a covariate and outcome. When multi-collinearity or other modelling assumptions necessary for inference are not accounted for, logical fallacies occur due to the tool. This all revolves around the idea of "fitness for purpose" and communicating the limits of methods and data and their likely best uses (Leyk *et al.*, 2019). More formally, this is a question of the epistemological fitness of a specific modelling framework for the question being asked. The question of whether these BSGM frameworks are best suited for causal inferential purposes, regardless of configuration, remains and should be compared against other potential frameworks. Within the BSGM frameworks, if one is using modelled disaggregative BS populations, created with covariates also used in determining BS transitions, the outputs of the BSGM modelling frameworks are more appropriate for end-use tasks such as further predictive modelling or more direct non-inferential applications (e.g. used as a spatial aggregating filter for healthcare access data).

The work in this thesis can look at and quantify patterns of BS change, but, themselves, can provide little insight into the causes of change, only the correlates of change. However, the framework has potential for producing outputs suitable for inferential end uses. Because the framework takes tabular estimates of BS population at some subnational level, the user is free to provide these estimates as produced by some means other than disaggregative population modelling. Additionally, the BSGM frameworks allow for covariates to be removed or added should a covariate normally used in the disaggregative modelling or in the spatial allocation portion of the BSGM frameworks be of causal interest. Here, I utilised disaggregative population modelling due to the absence of globally available subnational BS population estimates and corresponding BS extent data. In Chapter 5, the BSGM models were run with a reduced set of covariates limited to BS extent-derived covariates, topographic covariates, and information on protected areas. These reduced covariate versions are also the BS datasets that are publicly distributed, leaving as many possible end-uses available.

Here, I chose to create annual gridded population estimates and estimated BS extents as opposed to, say, seasonal or day/night largely because of the global coverage of this work and the resulting data availability. Official estimates of seasonal or day/night populations at the subnational level simply do not currently exist with global coverage. There is potential that novel datasets from Facebook or another social media platform (regarding app user location data) or Call Data Records could be used in the future, but this entails cooperation from numerous (if not hundreds of) companies, has ethical concerns regarding privacy, and a single device/login does not equate to a single person (Steele *et al.*, 2017; Weber *et al.*, 2018). Meanwhile, the Gridded Population of the World, version 4 (GPWv4) is the most complete subnational population count database with global coverage and it only provides annual estimates of population (Doxsey-Whitfield *et al.*, 2015). Further, high resolution (<= 100m resolution) maps of BS extent with global coverage were available at only at a few, i.e. five cross-sectional years, with only 3 of those years being within my study period of 2000 to 2020 (Pesaresi *et al.*, 2013, 2016; Esch *et al.*, 2013). No sub-annual BS extents estimates exist with global coverage. Given that I was interpolating between these 3 BS years, with an average of 7 years between them, to interpolate finer than annual would be implying a precision that the existing data could not support. With any dasymetric modelling technique there is a balancing act between the input and output spatial/temporal resolution of the modelling. That is, the greater the spatial or temporal scale difference between the source unit and the target unit, the greater the uncertainty in the disaggregation (Mennis and Hultgren, 2006; Schroeder, 2007; Nagle *et al.*, 2014; Zoraghein and Leyk, 2019). Temporally, I was disaggregating from periods of between 2 to 12 years down to single years. Spatially, I was disaggregating from subnational areas, with sizes ranging from a couple 100m^2 tens of thousands of 100m^2, to 100m^2 pixels. So, in summary, I didn't opt for producing sub-annual estimates of BS nor population given that the available data, e.g. large gaps between observations and lack of sub-annual specific data that was globally available and comparable, and prioritising creating a flexible and automatable modelling method as well as data output that was globally applicable and comparable. As more data becomes available, refinements to the temporal and spatial scale of predictions can be explored.

The findings of this thesis are limited to the chosen BS data representations utilised here. There are different built-environment and BS data sets with differing radiometric and operational definitions, especially if we open the scope to non-global datasets, with all covering some aspects of the physical component of urban. As computational resources and software become more available and accessible, these types of datasets are becoming more and more available and their characteristics, definitions of the built-environment, accuracy, and quality become more varied. How generalisable the BSGM framework is to this variety of definitions is largely unknown. However, in this thesis, I have utilised three of the most prominent and contemporary global BS datasets (ESA Annual, GHSL, and GUF) with consistent results.

The validations of Chapter 3 and Chapter 4 were done across entire countries. While the unit of measures were either pixels or subnational, there was variation in the model performance. While detailed examination of the validations within subregions of countries have not been done, visual examinations give some indications as to how the models perform within countries. Much like the input BS data, the model frameworks have a bias near established BS agglomerations, i.e. in urban and peri-urban areas. Consequently, the model frameworks are more likely to display infill and expansive type growths as opposed to "leapfrog" or spontaneous type growth. This is not to say that the model is incapable of spontaneous growth scenarios. When the subnational units are relatively moderate to fine in spatial resolution and lights at night data are available, spontaneous growth does become modelled. However, if the input BS data does not capture, say, a small settlement, then the BSGMi, due to its interpolative nature, cannot predict the small settlement occurring. However, when extrapolating, the annual weighting of the RF-derived period transition probabilities by the time lagged lights at night data, e.g. a scenario where it is 2020, you are predicting forward from 2015 and you have lights data through 2019, becomes very important for near future spontaneous settlement prediction.

Other ways to address the current urban-centre bias in the BSGM frameworks could take several forms. While the current framework models each subnational unit independently, the “Demand Quantification” component could be reworked to utilise a spatial hierarchical modelling framework, e.g. using the Integrated Nested Laplace Approximation (INLA) package (Illian, Sørbye and Rue, 2012). This would have the drawback of added computational time and more manual model fitting, but the information contained in the spatial arrangement of the subnational units could be leveraged and subnational units could even be classified as urban, rural, and peri-urban for further differentiation. Another potential is to classify the units *a priori* based upon their urbanity and stratify the BSGM framework to work independently on each class of units. Of course, deciding on an applicable classification scheme, especially one that is globally applicable, could be difficult. These potential modifications would also lend themselves to providing more informed future projections as well.

A further limit is that this model only produces short-term, i.e. up to five years, future predictions of BS extents. This is partially due to the generalised guidance on ARIMA and ETS model use (Chapter 2, Section 2.3.2.1), but more so due to the short time-series of data I had for BS extents. Referring to the GHSL-GUF BS representation partially discussed in Chapter 5, I had only 3 years of observed data: 2000, 2012, and 2014. Using the BSGMi, I was able to fill in the temporal gaps between these observations to have annual estimates of BS extent from 2000-2014, resulting in a time series of 15 years. However, even though validation, using a different BS dataset (Chapter 3), showed the BSGMi had the potential to be highly accurate, the error of the BSGMi when using the GHSL-GUF data was unknown due to the small sample precluding withholding data for validation. If substantial errors exist in the interpolated BS extents, then this error would continue to propagate and grow the further the BSGMe would extrapolate. Even assuming to have 15 years of observed data, 15 observations is a very small sample for time-series methods like ARIMA and ETS models. Therefore, limiting the end use to near-future projections can serve as a simple conservative limit on the error in the predicted future extents, as demonstrated in the findings of Chapter 4.

There is potential for longer term future projections, e.g. 20, 30, 50-years, of BS expansion using the BSGMe framework. The best conditions for this to occur would be in the scenario where population data and the corresponding subnational units are relatively fine in spatial resolution and there are numerous time points upon which to fit the logistic growth/decay curves and the natural cubic splines. Ideally, the projected population numbers would be independently derived from demographic models (Booth, 2006; Hyndman and Booth, 2008), utilising fertility and mortality as inputs as opposed to the simple, yet global in coverage, exponentially extrapolated population data used here (Doxsey-Whitfield *et al.*, 2015). Many of these demographic models produce “scenarios” with “low” to “high” population growth that could facilitate the production of corresponding low-to-high BSGMe predictions, similar to contemporary long term urban growth forecasts (Gao and O’Neill, 2019, 2020). Remaining with the current BSGMe framework, the predictions are relatively simplistic. While this allows great flexibility and minimal data requirements, it could result in less than nuanced long-term predictions. For instance, the performance bias of predictions near established BS agglomerations also means that predicting for small settlements are relatively poorer, especially in data poor contexts, e.g. few and large subnational units or few timepoints informing the extrapolation. Further, it is still unknown how well the future predictions do or do not capture the origination of new, isolated settlements. If the BSGMe were to be further developed for longer-term predictions a separate module, in addition to the existing Demand Quantification and Spatial Allocation modules, would need to be developed. Such a module would predict the number of new isolated settlements occurring, possibly as a Poisson point process (Diggle, 2014). It would also have a separate suitability layer determining where these new, isolated settlements should occur. Alternatively, land-use transition models can provide a more nuanced simulation of land use change. This class of models look at land use interaction, land use conversion costs, and other drivers of land change conversion, but they require more user input, expert defined parameters, and have higher base data requirements (Verburg *et al.*, 2002; Verburg and Overmars, 2009; Schaldach *et al.*, 2011; van Asselen and Verburg, 2013; van Vliet, Eitelberg and Verburg, 2017).

These points being said, there are no mathematical reasons that the current BSGMe forecasting methods cannot predict subnational BS demand cannot predict indefinitely into the future. The forecasting methods estimating the magnitude and timing of the BS growth at the subnational level, i.e. the ARIMA, ETS, and GLM models within the “Demand Quantification” component of the BSGMe (Chapter 4), are strictly autocorrelative. That is, they are not dependent upon any external covariate values and only rely on their previous values to predict their future values. Of course, as with any extrapolative prediction, the prediction uncertainty increases the further one predicts from the last observation.

The RF models within the "Spatial Allocation" component of the BSGMe, which are used in partially determining where the predicted subnational level BS growth is distributed to pixels within the subnational unit, and the framework surrounding them limit future predictions much more. This is in part due to data constraints, such as covariates input into the RF not being available or easily modelled into the future, and partially due to the way I implemented the RF.

For discussion, let us assume the RF was made using a minimum set of covariates including distance to nearest BS edge, proportion BS within several radii, elevation, and slope. These variables are either capable of being modelled by the BSGMe or assumed time invariant (elevation and slope). Let us also assume that we have a RF trained to predict the probability of non-BS-to-BS transitions between 2010 and 2020 and that we are interested in predicting BS extents from 2020 to 2030. Giving the RF covariates corresponding to 2020, we make predictions of the pixel level probability of non-BS-to-BS transition occurring. However, this probability is based upon the assumption that the 2010 to 2020 period's relationships between the covariates and the probability of transition remain the same for the 2020 to 2030 period. Further, because the RF was trained to predict the probability of transition over a ten-year period, it is only appropriate to predict for another 10-year period. For instance, if we were interested in predicting across the period of 2020 to 2035, a RF trained over a 10-year period would not be providing 15-year probabilities. In this scenario, the 10-year RF would likely be underestimating the probabilities of transitioning, due to less years or "opportunities" of transitioning. Conversely, if a 10-year RF were used to predict over a 5-year period, the probabilities would likely be overestimated.

If the BSGMe framework were to be applied to produce more medium and long-term estimates of future BS change, e.g. 15, 20, or 50 years, great consideration would need to be given to the way the RFs were constructed and applied. I would speculate that having many RFs covering smaller periods would be the most useful scenario, but this would require an expansion of the coverage and periodicity of the input BS time series of data, e.g. coverage over a greater period of time and BS extents every 5 years as opposed to, say, 10 years. That is, because of RFs ensemble nature several small period RFs can be naturally combined to produce a RF representative of a larger period. Alternatively, a small period RF could be reused iteratively. For example, assuming we are still using the minimum set of covariates, a 5-year RF (trained from 2015 to 2020) can be used within the BSGMe framework to produce predictions of BS extent across

2020 to 2025. The predicted 2025 BS extents can be used to derive new BS covariates to input back into the same 5-year RF to then produce predictions from 2025 to 2030 and so on. This is, of course, still predicated on the assumption that the relationships between the covariates and the probability of non-BS-to-BS transition remain static; something that still requires exploration.

6.2 Future Work

Having expanded the globally available time series of BS extent data, larger scientific and methodological questions can begin to be addressed or addressed with greater detail and in a globally consistent and comparable manner. For instance, an expanded consistent time series of BS extents could be used to investigate temporal relationships to other human influenced land use patterns, e.g. rates of conversion for various land cover classes to BS. Similarly, identifying rates, locations, and spatial patterns of BS expansion could allow for better classifying types of settled areas, types of general urban expansion, and exploring the implications for sustainability of such expansion. Such implications that could be investigated include water and food security, loss of arable land, and increased human-wildlife interaction, a potential vector for novel diseases. Additionally, this time series of consistent binary BS extents could be adopted under an urban definitional framework, e.g. the REGIO model (Dijkstra and Poelman, 2014) and integrated with other data for creating a time series of broader urban classification(s). This could then be used for describing changes in urban intensity over time and space as well as examining their relationship to environmental impacts, such as the emission of greenhouse gases.

This type of globally consistent, comparable, and applicable BS modelling framework also has ready applications in other recent population and urban research efforts, such as the GRID3 project (https://www.grid3.org) and the GHS-Settlement Model (GHS-SMOD; https://ghsl.jrc.ec.europa.eu/ghs_smod2019.php) projects.

GRID3 is carrying out bottom-up population mapping and is heavily dependent on survey data and knowing where settlements are. Settlement data is being provided to them by commercial data providers at one or two time points, which may or may not temporally align with the survey data. Or the settlement data may line up temporally with the survey data, but forward or back projection of the spatial population distributions are needed. GRID3 is also carrying out top-down population modelling where which is also strongly informed by settlement related data (Nieves *et al.*, 2017; Stevens *et al.*, 2020; Reed *et al.*, 2018). The BSGM frameworks can expand the temporal coverage of the settlement data they are provided without sacrificing the spatial resolution and maintaining the radiometric definition of the input settlement data. This in turn facilitates population modelling and or projections while maintaining a comparable and consistent BS definition, capturing small settlements, and growing settlement with population.

GHS-SMOD is a framework that uses the Global Human Settlement Layer (GHSL) extents, gridded population, and the Degree of Urbanisation (DEGURBA) framework (European Commission and Statistical Office of the European Union, 2019) to define various degrees of urban intensity across the globe using a comparable definition. However, GHSL only has 4 time points of coverage in 1975, 1990, 2000, and 2014. The BSGM could be utilised to interpolate between these years and further facilitate the application of the GHS-SMOD framework to the interpolated years. This would provide greater opportunity to examine how urban areas under the GHS-SMOD framework have evolved over time and how these evolutions have correlated with other land use, economic, and population drivers. And insights into the drivers and correlates of settlement and urban expansion gathered from GHS-SMOD could stratify and refine the BSGM modelling approaches.

On a more technical note, future explorations of temporally extending and validating the BSGMe framework are a logical next step. As the time series of globally available BS extents become longer, providing more input to the modelling framework, BS extent predictions would be assumed to become more accurate and or less uncertain. At the very least, such an increase in the BS extent time-series data would allow for larger training sets. The BSGM framework can provide a foundation for translating this increase in information into more medium- and long-term BS extent forecasts. Similarly, repurposing the BSGMe to extrapolate backwards in time from a last observation would be useful, for example in extending estimated BS extents to pre-Landsat (pre-1975) coverage, and should be investigated. Overall, future technical work related to built-settlement expansion modelling should focus on five key areas:

- Extending the future and past projections as new time series of BS extents become available

- Further exploration of the limits of using relative population change as an indicator of non-BS-to-BS transitions at finer spatial scales
- More explicit, and potentially time specific, incorporation of transportation networks in the modelling framework and process
- More explicit incorporation of the spatial arrangement of the data, e.g. a spatial weights matrix, in the modelling process thereby leveraging information from the "neighbourhood",
- Attempts to better quantify the uncertainty in the model outputs either through a different modelling approach or maps providing quality assessments of the input data and,
- Exploring hierarchical or stratified approaches to modelling wherein different subnational areas and or settlements are grouped or classified and modelled accordingly

More broadly, the expanded and consistent global BS datasets produced under this framework provide future opportunities to address larger scientific questions. These expanded datasets can be used to examine the trends in settlement expansion in low lying coastal regions, the impacts of settlement expansion on air pollution and greenhouse gas emissions, and the impacts of settlement expansion and exposure to vector borne diseases. Having a consistent time-series of BS extents allows for the examination of how human settlement evolve over space and time, the correlations with migration, various scales of economy, and in relation to conflicts. By examining the spatial and temporal variations of these BS data, better guidance on balancing policy recommendations can be developed across scales, from global to national to local.

Future work should also investigate the effect of scale regarding the assumptions of the BSGM frameworks while maintaining a consistent RS-based definition of BS. In the supplemental material of paper 1 (Appendix A), I showed that using a fine scale BS representation, GUF Evolution with an original resolution of 72m performed poorly in validation, relative to the 300m original resolution of the ESA Annual data. Direct comparison between the GUF Evolution and the ESA Annual BSGM model performances was not possible due to differences in model extent and BS definitions of the datasets. However, whether the assumptions of the BSGM modelling framework hold when using finer scale and/or different definitions of BS is an interesting one with, currently, mixed evidence. **Error! Reference source not found.** demonstrated that using a combination of the GHSL (38m native resolution) and GUF (72m native resolution) to produce BSGM-derived predictions were important for predicting population density at the subnational level. However, Tayyebi et al. (2013) investigated modelling urban growth in the USA and found that at finer spatial resolutions, accuracy of predictions decreased.

Given that urban accessibility, a covariate derived from road network data, was found to be one of the most important covariates in predicting where BS expansion would be spatially located (Chapter 3), future work should look at improving road/transportation data and connectivity of settlements. Time-specific road and transportation networks would likely be invaluable in predicting the location of BS expansion.

To quote the band Arcade Fire's song "Wasted Hours," "First they built the road, then they built the town." While Arcade Fire was describing North American suburban expansion in the late 20th century, there is some generalisability to this statement. Any development of structures or settlement will require transportation of goods to the development site, which typically requires transportation infrastructure in the shape of roads or paths. If such infrastructure change can be captured, then new development and BS expansion may be better identified and is certainly worth further investigation.

At another scale, utilising the contemporary road data could be used to assess the connectivity of settlements and the spatial configuration of that network. In conjunction with demographic and other BS data, this would allow for the establishment of typologies of settlements based upon their spatial arrangement, spatial linkages, and characteristics. These typologies could then be used to better understand BS expansion and or intensification, economic developments, demographic transitions, and disease transmission, to name a few applications.

The modelling framework I presented here treats every subnational unit independently. Future work could explore the utilisation of spatial weights, e.g. as in a geographically weighted regression, to leverage information from neighbouring subnational units in the modelling process. This could help improve predictions in units with little or noisy information and smooth out the predictions. If settlements were able to be classified, e.g. by population or area size or functional purpose, or grouped in a hierarchical manner, stratified or hierarchical modelling approaches could be utilised to "share" information across the model.

Although datasets of contemporary and past measures of the extents of urbanised areas, the built-environment, and settlements continue to grow at a rapid rate, the demand for predictions of these phenomena into the future will remain. As long as that demand exists, models capturing the spatio-temporal trends of these extents will have predictive and explanatory utility and should similarly grow and evolve.

Appendix A Supplemental Document for Chapter 3

Section A1 – Covariates Used in Population Map Creation and Their Sources

Table A1. Covariates utilised in the production of the population maps that were used as inputs into the built-settlement growth model

Covariate	Time Point(s)[a]	Original Source	Source Resolution
DTE Cultivated landcover	2000, 2005,2010, 2015	ESA CCI Landcover (ESA CCI, 2017) classes 10-30	10 arc seconds
DTE Woody, Herbaceous, Shrub landcover	2000, 2005,2010, 2015	ESA CCI Landcover (ESA CCI, 2017) classes 40-120	10 arc seconds
DTE Grassland landcover	2000, 2005,2010, 2015	ESA CCI Landcover (ESA CCI, 2017) class 130	10 arc seconds
DTE Lichens and Mosses landcover	2000, 2005,2010, 2015	ESA CCI Landcover (ESA CCI, 2017) class 140	10 arc seconds
DTE Sparse Vegetation landcover	2000, 2005,2010, 2015	ESA CCI Landcover (ESA CCI, 2017) classes 150-153	10 arc seconds
DTE Aquatic Vegetation landcover	2000, 2005,2010, 2015	ESA CCI Landcover (ESA CCI, 2017) classes 160 - 180	10 arc seconds
DTE Bare Areas	2000, 2005,2010, 2015	ESA CCI Landcover (ESA CCI, 2017) class 200	10 arc seconds
DTE Built-settlement	2000, 2005,2010, 2015	ESA CCI Landcover (ESA CCI, 2017) class 190	
Distance to Inland Water Bodies	2015, assumed invariant	MERIS-based water bodies (Lamarche et al., 2017)	5 arc seconds
Distance to Roads	Downloaded 2017, assumed invariant as temporally specific road data unavailable	(OpenStreetMap Contributers, 2017) (OpenStreetMap Contributers, 2017)	Vector
Distance to Rivers	Downloaded 2017, assumed invariant	(OpenStreetMap Contributers, 2017)	Vector
Distance to Coastline	Based upon boundaries of GPWv4, assumed invariant	CIESIN GPWv4 (Doxsey-Whitfield *et al.*, 2015)	Vector
Slope	2000, assumed invariant	World Wildlife Fund Voidfilled Hydrosheds (Lehner, Verdin and Jarvis, 2008)	3 arc seconds
Elevation	2000, assumed invariant	World Wildlife Fund Voidfilled Hydrosheds (Lehner, Verdin and Jarvis, 2008)	3 arc seconds

DTE: Distance To nearest Edge

a Note, for any covariate derived from land cover or built-settlement, only one year-specific covariate was used corresponding to the desired population surface (e.g., for a 2000 population surface only covariates corresponding to 2000, or those assumed temporally invariant, were used as covariates).

For every population year modelled, we included the distance to nearest BS edge for the year 2000, as population relates to older parts of a BS agglomeration differently from younger ones (Andrea E. Gaughan *et al.*, 2016). 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 (Andrea E. Gaughan *et al.*, 2016).

Section A2 – Full Process Diagram and Additional Rationale

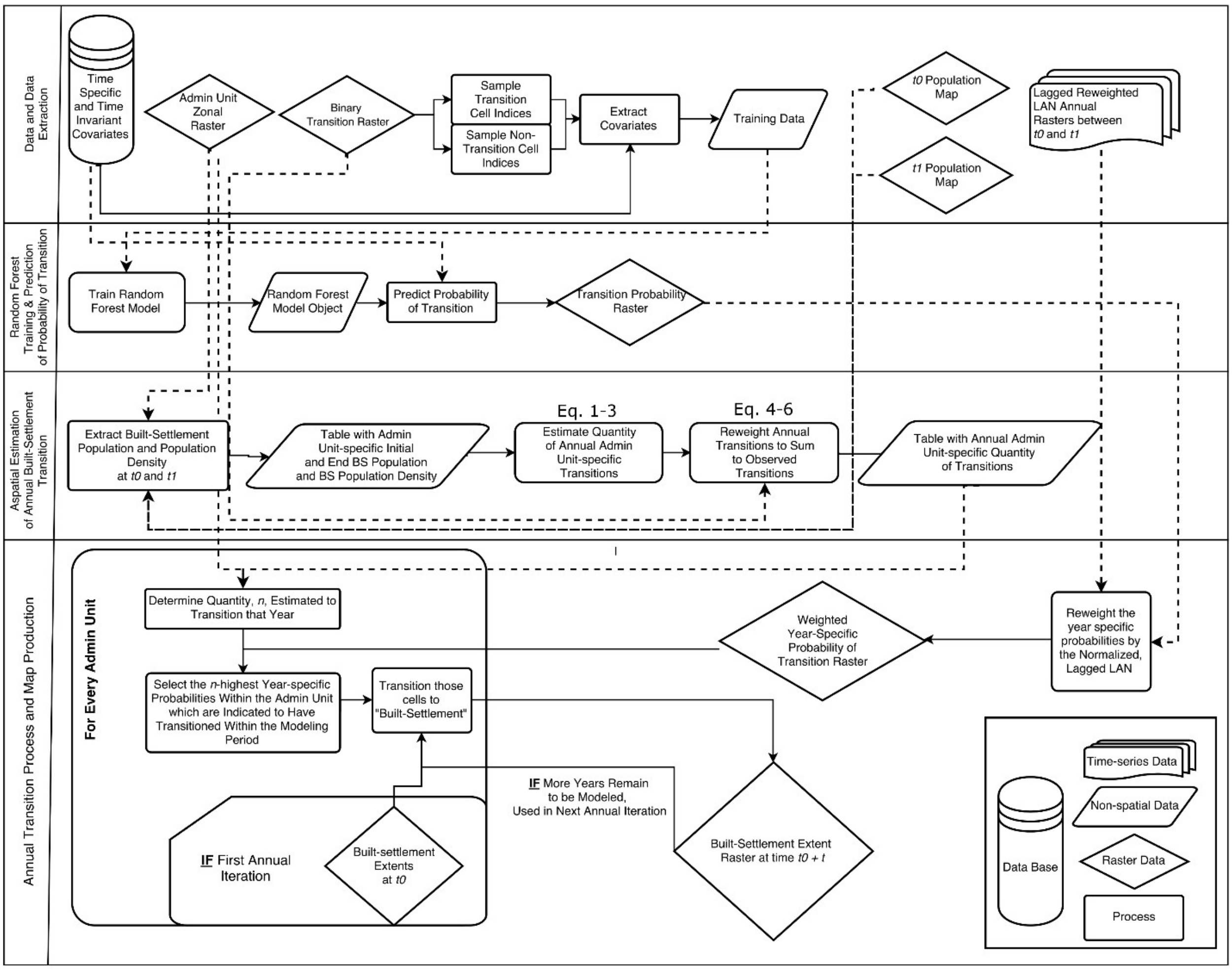

Data and Data Extraction
Time Specific and Time Invariant Covariates
Admin Unit Zonal Raster
Binary Transition Raster
Sample Transition Cell Indices
Sample Non-Transition Cell Indices
Extract Covariates
Training Data
t0 Population Map
t1 Population Map
Lagged Reweighted LAN Annual Rasters between t0 and t1
Random Forest Training & Prediction of Probability of Transition
Train Random Forest Model
Random Forest Model Object
Predict Probability of Transition
Transition Probability Raster
Aspatial Estimation of Annual Built-Settlement Transition
Extract Built-Settlement Population and Population Density at t0 and t1
Table with Admin Unit-specific Initial and End BS Population and BS Population Density
Eq. 1-3
Estimate Quantity of Annual Admin Unit-specific Transitions
Eq. 4-6
Reweight Annual Transitions to Sum to Observed Transitions
Table with Annual Admin Unit-specific Quantity of Transitions
Annual Transition Process and Map Production
For Every Admin Unit
Determine Quantity, n, Estimated to Transition that Year
Select the n-highest Year-specific Probabilities Within the Admin Unit which are Indicated to Have Transitioned Within the Modeling Period
Transition those cells to "Built-Settlement"
Weighted Year-Specific Probability of Transition Raster
Reweight the year specific probabilities by the Normalized, Lagged LAN
IF More Years Remain to be Modeled, Used in Next Annual Iteration
IF First Annual Iteration
Built-settlement Extents at t0
Built-Settlement Extent Raster at time t0 + t
Time-series Data
Non-spatial Data
Data Base
Raster Data
Process

Figure A1. Overview of the generalised modelling process for a case of only two observed timepoints, *t0* and *t1*, with references to utilised equations.

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.

Logistic growth curves are widely used and accepted for modelling populations within demography, ecology, and urban modelling (Austin and Brewer, 1971; Wilson, 1976; Ledent, 1982; Cohen, 1995; Smith, 1997). Batty (2009) summarised, “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 *et al*. (2005) 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 Verhulst (1838). Furthermore, Ledent (1982) showed that urbanisation, the process of *population* becoming urban, across time can be adequately summarised by “S-shaped curves”, specifically the functional logistic form. We followed this same underlying conceptual logic using a 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.

For each unit, we interpolated the corresponding unit-average BS population densities, referring to the years t in T, across all unobserved years t_k. Because there is a lack of literature and data on the actual or theoretical limits of human population density, we selected natural cubic splines to interpolate each unit’s BS population density while avoiding the sharp rates of change, that would be seen with piece-wise linear interpolation, or large oscillations seen with higher order polynomials (i.e. Runge’s phenomenon). Cubic splines have a long history in demographic interpolation, including interpolation of rates associated with urbanisation processes (McNeil, Trussell and Turner, 1977; Ledent, 1982). Here 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.

Section A3 - Handling of Negative or “Decay” Transition Cases

The process resulting from Equations 1-3 can produce "negative" predicted growth, hereafter decay, in any given year, and that the input built-settlement extent data assumes that once an area has transitioned to built-settlement it remains built-settlement. To account for this, we used the input data to limit the model to show "stagnation," i.e. no growth, or growth. We managed three case types of decay according to the information presented in Table A3. Case I (A and B) included situations where the observed transitions were greater than zero and some or all estimated transitions were negative. Case IIs included situations where the observed extent transitions were zero and the estimated transitions were negative. For full details, read comments in model code.

Table A2. Case types of predicted built-settlement decay and how they were handled in the model.

Case Type	Sub type	Description	Origin	Implication	Handling
I	A	All predicted years < 0; observations > 0	Built population decreases, built settlement increases because of differences in imagery and sensitivity of original datasets or because relationship between population and built settlement area are inverse of what would be expected.	Lacking any other information, we will assume that the greatest built settlement changes occurred circa the biggest population magnitude changes	The reweighting scheme makes all the weights positive by virtue of all individual differences being negative; no special action is necessary.
	B	Some predicted years < 0; observations > 0	Comes about from population decreasing for a year while outpacing the predicted decrease in built settlement population density	Built settlement growth during this period is unlikely compared to other years in the total transition period.	Set the predicted difference for that year, and therefore its weighted difference, to zero.
II	---	Some, but not all, predictions < 0; Observations = 0	Relationships between population and built settlement counts are not straightforward and not necessarily stationary through time and or space. Further, inaccuracies exist in the original built settlement data and the popA6ulation estimates. Any of these errors, in conjunction with model assumptions, could combine to result in this.	Continue with the base assumption that the input built-settlement data is the best we have in knowing if any transition occurred.	Set all predicted differences to zero in order to match the observed changes

Section A4 – Stochastic Process for Obtaining Agreement After Rounding of Predicted Transitions

After the negative values were handled (See Supplemental Material, section A3) and the observed transitions were dasymetrically redistributed, sometimes there remained discrepancies, i.e. over or under estimations, between the sum of predicted transitions and the observed changes due to rounding during the weighting procedure. Here, we obtained agreement between the predicted and observed transitions by way of a stochastic process where, as long as the predicted number of transitions of a given administrative unit did not equal the corresponding observed transitions, we randomly selected a time point within the modelling period. We then added or subtracted, whichever was appropriate, one transition to the total predicted transitions for that year, $BSCNT_{ti}$. This "salting" continued until agreement between the predicted and observed counts for a given admin unit was obtained. When we performed subtraction to correct for overestimation, we did not subtract from years that we already predicted to have no transitions.

Section A5 – Additional Results Based Upon ESA Input to BSGM

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 A3). At worst, e.g. Vietnam 2002, approximately half of the units were still performing better than the naive model (Figure A3). For quantity disagreement (Figure A4) and allocation disagreement (Figure A5), results similar to pixel level results were found.

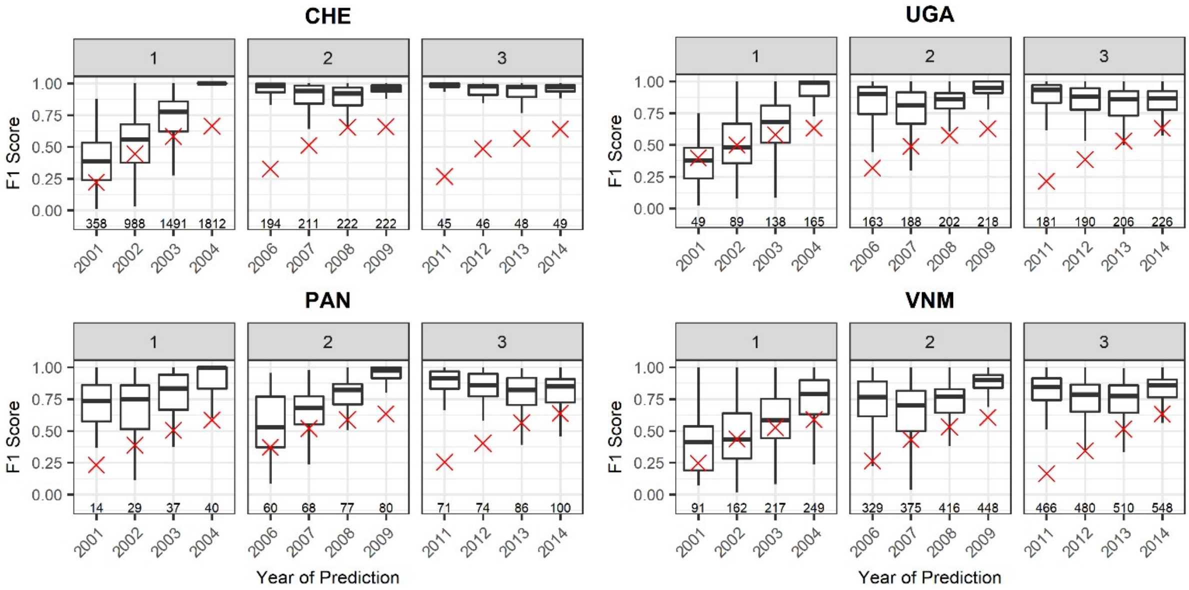


Figure A3. Unit level F_1 score box plots, by dasymetric period, of Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) ESA informed models as compared to a naive model, given by a red “x”. Number of units exhibiting any transitions for each period and a defined metric value is given above the x-axis.

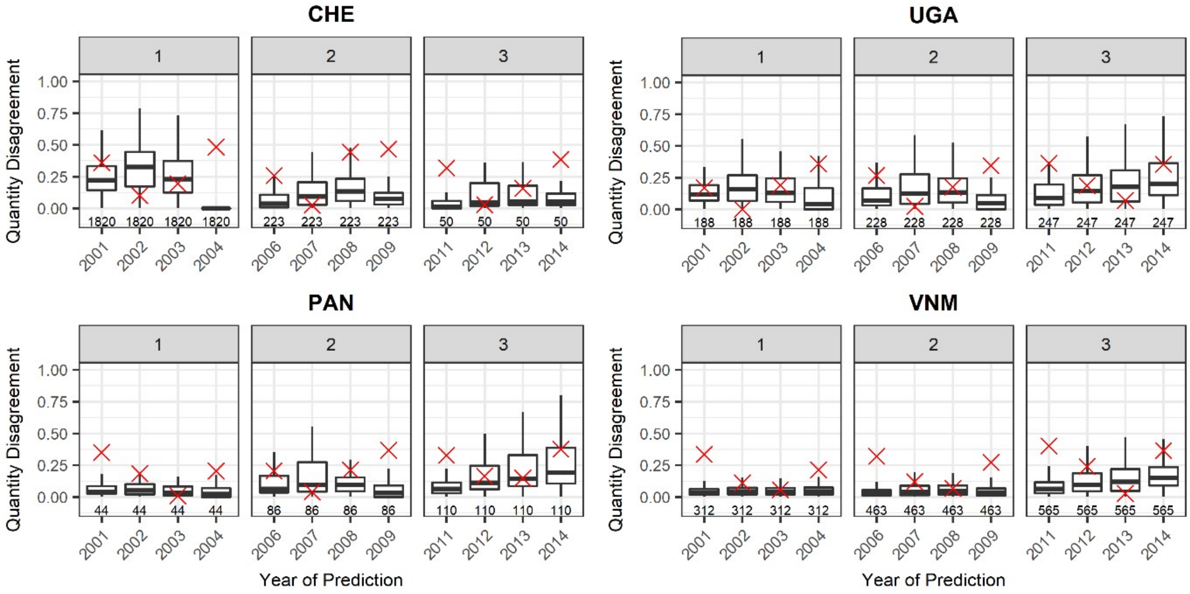


Figure A4. Unit level quantity disagreement box plots, by dasymetric period, of Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) ESA informed models as compared to a naive model, given by a red “x”. Number of units exhibiting any transitions for each period and a defined metric value is given above the x-axis.

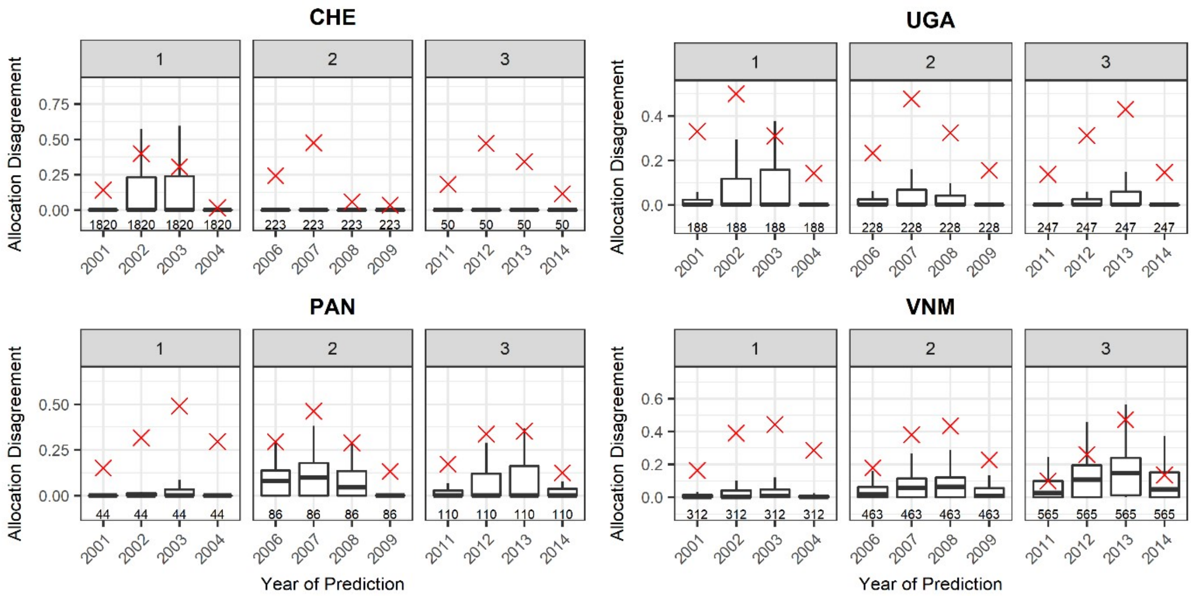


Figure A5. Unit level allocation disagreement box plots, by dasymetric period, of Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) ESA informed models as compared to a naive model, given by a red “x”. Number of units exhibiting any transitions for each period and a defined metric value 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.

Examining the year-specific study area F1 scores (Figure A6), we show that the BSGM modelling framework had low absolute performance and near naive model performance between 2001 and 2003 across all countries. After 2003, the F1 score notably, with values approaching 1.0 in some cases, and the BSGM modelling framework dramatically outperforms the naive model (Figure A6).

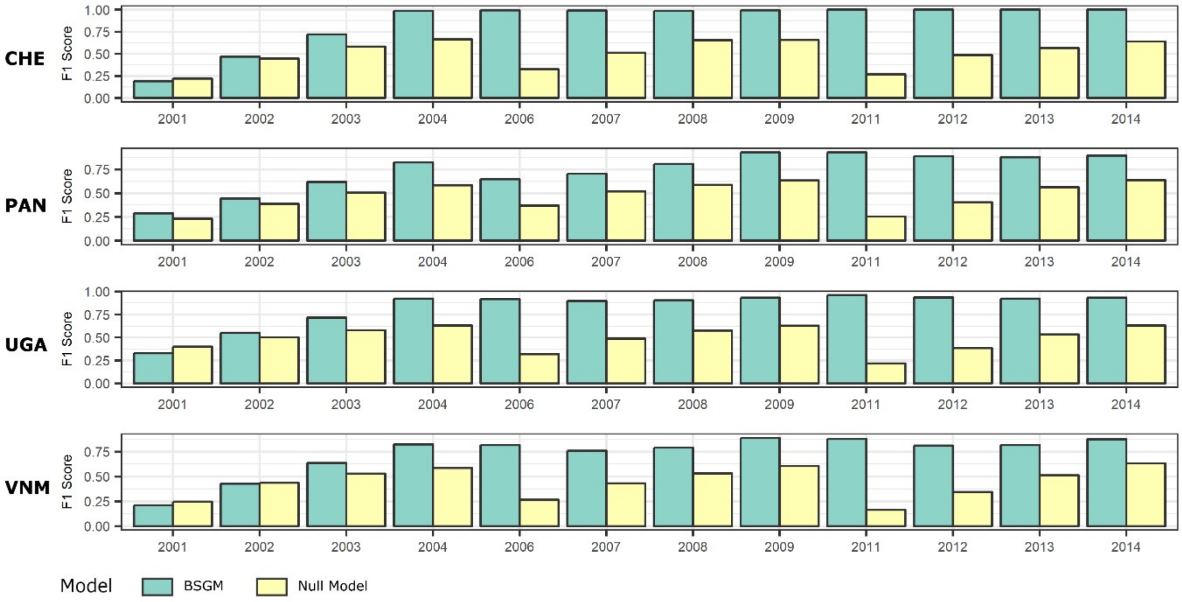


Figure A6. Pixel-level F_1 score by year for the BSGM-based BS extents and BS extents produced using the naïve model for Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) Full annual contingency data and metrics in supplementary material

Section A6 – BSGM Results with GUF Evolution

To demonstrate the flexibility of the modelling framework and its applicability to higher resolution, e.g. sub-100m, urban feature data, we also tested an alpha version, of the forthcoming World Settlement Footprint multi-temporal dataset, known as WSF Evolution (Esch 2018a), hereafter WSF Evo. Starting in 2018 with a release of the WSF 2015 (equivalent of binary GUF, based on a joint analysis of multi-temporal Sentinel-1 and Landsat-8 data for the year 2015), the WSF-Evo product provides detailed information about the spatio-temporal development from 1985–2015 for each human settlement identified in the WSF-2015. The corresponding analysis is based on a processing of multitemporal collections derived from the Landsat archive using an implementation of the TimeScan approach (Esch, 2018b) at the Google Earth Engine (Gorelick *et al.*, 2017) to generate the baseline layer for the classification. This classification starts by using the WSF 2015 as training data for the identification of the built-up area in 2010, and the using the resulting WSF 2010 as training input for classifying the 2005 data, and so on. The WSF Evo. Data used here covers a 50km x 50km rectangular area centred over Ho Chi Minh City for the dates of 2000, 2010, and 2015. The data was resampled to 3 arc seconds using nearest neighbor resampling and derived covariates were calculated from this. For the purposes of modelling, we only utilised areas that completely covered the subnational units in the population data. Summaries of the study areas with regards to BS transitions as defined by WSF Evo. are given in Table A3.

Table A3. Descriptive summary of the WSF Evolution dataset where areal units are pixels pixels (~100m) as that is the unit handled by the model which looks at relative areal changes as opposed to absolute areal changes.

Dataset	Country [a]	Period	Initial Non-Built Area (pixels)	Observed Transitions
WSF Evo.	Vietnam	2000-2015	3,295,142	10.43%

a Ho Chi Minh City and immediate surroundings

The RF using WSF Evo. Data, out of all the models, has the largest area under its PRC curve, but the precision begins to decrease, albeit less sharply, at lower recall levels than the ESA models (Figure A7).

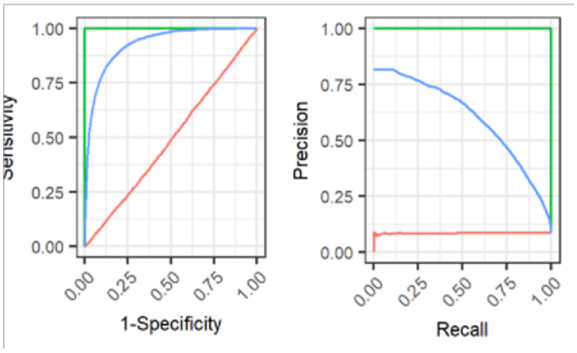


Figure A7 Receiver Operator Curve (left plots) and Precision Recall Curves (right plots) with the RF model performance, blue lines, against a random model, red lines, and a perfect model, green lines, for each modelled country and input dataset

Covariate importances for the WSF Evo. (listed as GUF or GUF+ in Figure A8) model were comparable to that seen in the ESA models (Figure A8).

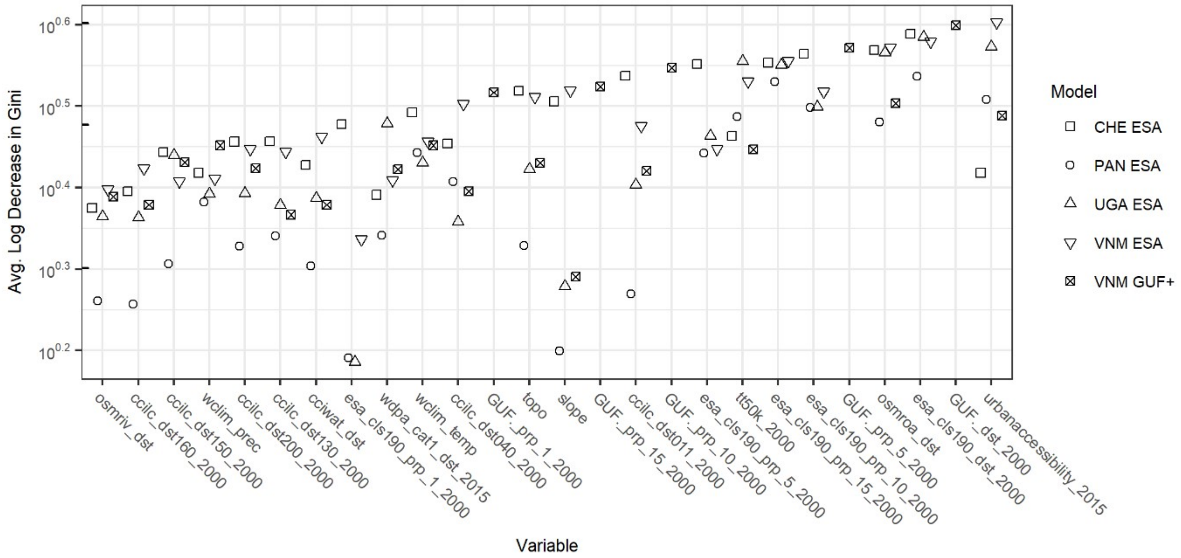


Figure A8. Random forest covariate importance as measured by the average log decrease in the Gini impurity when the covariate is used as the splitting criteria at nodes; higher values indicate better performance of covariate. Model for Swizerland (CHE) ESA, Panama (PAN) ESA, Uganda (UGA) ESA, Vietnam (VNM) ESA, and Vietnam WSF Evo. (GUF+) are shown. Refer to Table 1 for covariate names.

Overall, at the pixel level WSF Evo. Model performed much poorer than the ESA models with an overall accuracy of 0.518 (Table A4).

Table A4. 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.

Model	2001	2002	2003	2004	2006	2007	2008	2009	2010	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 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
VNM WSF Evo.	---	---	---	---	---	---	---	---	0.518	---	---	---	---

For the single compared year modelled using WSF Evo. data, the F1 score performance is low both in absolute terms, approximately 0.33, and in relative terms, having a score approximately 0.05 higher than the null model (Figure A9).

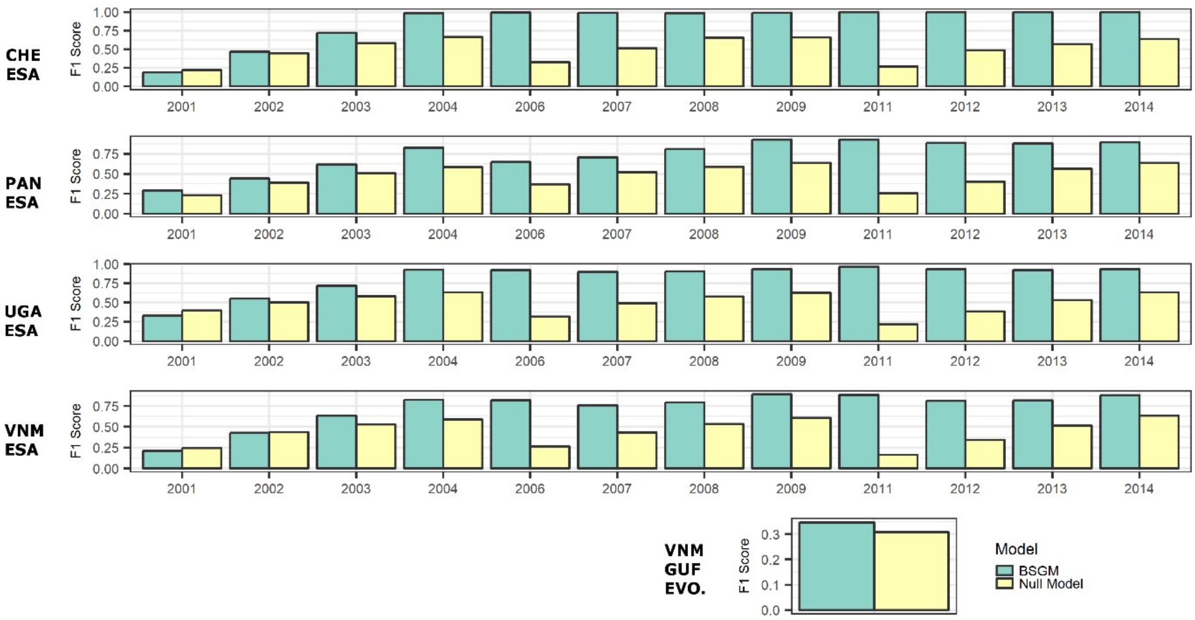


Figure A9. Pixel-level F$_1$ score by year for Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) as compared to a null model.

For the year modelled using WSF Evo., the total disagreement due to the BSGM is only slightly less than the null model primarily due to less allocation error (Figure A10).

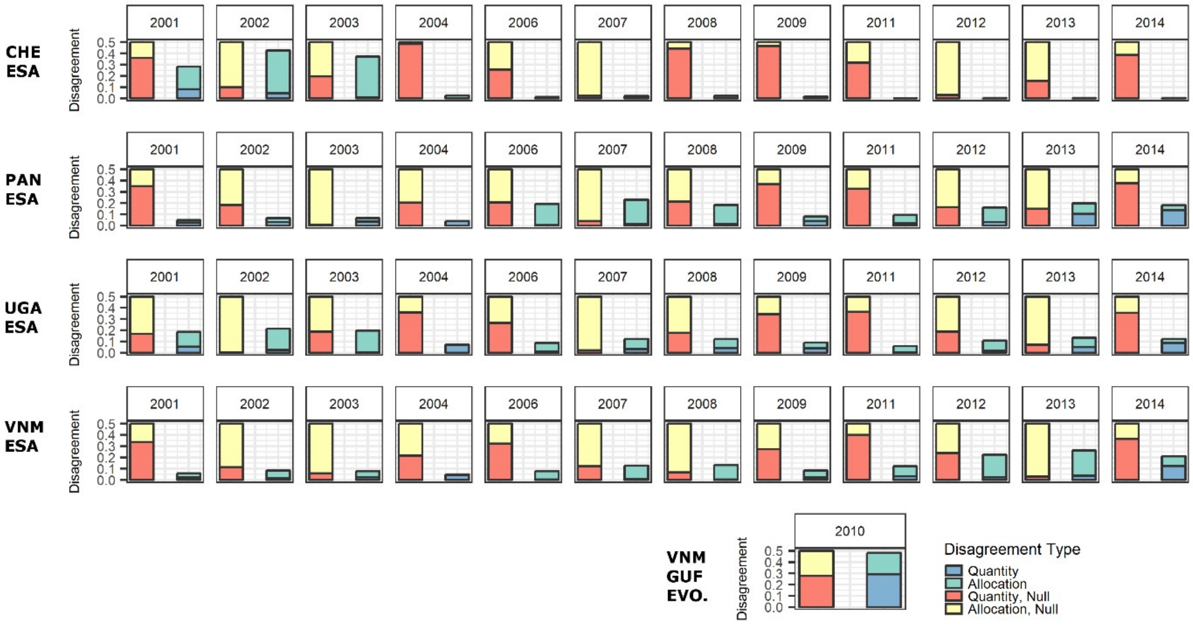


Figure A10. Pixel-level quantity and allocation disagreement of BSGM and null models for Switzerland (CHE), Panama (PAN), Uganda (UGA), and Vietnam (VNM) as compared to a null model, given in red. Full annual contingency data and metrics in supplementary material

For the WSF Evo. model, most of the units perform better than the null model in terms of F1 score and allocation disagreement (Figure A11). Conversely, many units have higher quantity disagreement than the null model, pointing to the underlying model performance issue being related to the population and population density interpolation as related to the assumed relationships with the input BS data rather than the RF and LAN allocation portion of the model.

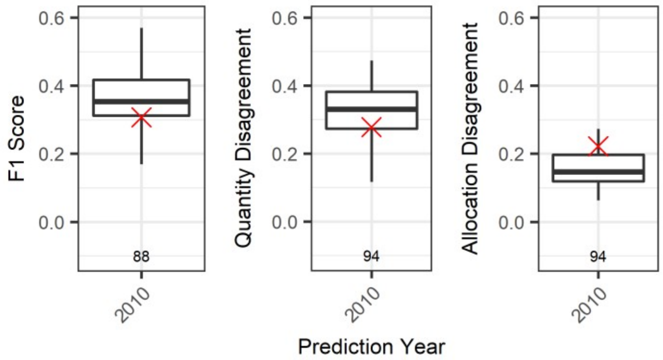


Figure A11. Using WSF Evo data, unit level distributions of F_1 score, quantity disagreement, and allocation disagreement.

The poor performance of the BSGM with the WSF Evo. data (Figures A9-11; Table A4) was surprising considering the excellent performance with the ESA-informed models (Table A4; Figures A9-10). There could be many reasons for this, but we believe it originates because of a dissonance between the assumed population growth relationships and the input data in the form of one or the more of the following:

i) temporally differing biases in the modelled RF-informed population surface

ii) less observed points to base the model upon

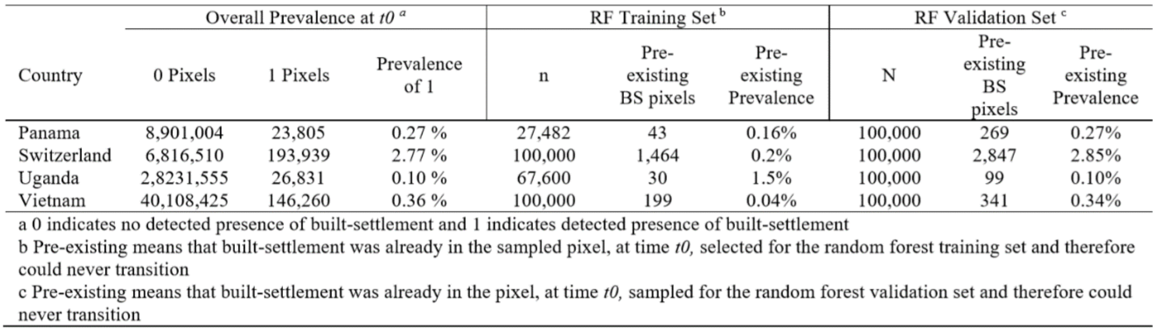
iii) given the experimental nature of the WSF Evo. data, it is unvalidated and therefore how well it is capturing the BS extents at any given time point is largely unknown (but assumed to be better than the ESA class 190)

iv) the WSF Evo data is capturing things perfectly, but the relationships the model can currently capture and describe are not sufficient to match the relationships between this spatial scale of data and the phenomenon.

More sensitivity analyses and validation testing is needed with finer scale input data across a larger time period of observations and a nested model, with varying behaviour for small type BS agglomerations and large type agglomerations of BS, should be carried out.

Section A7 –Training and Validation Sets of the Random Forest

Table A5. Descriptive statistics of pre-existing prevalence of BS in the training and validation datasets used for measuring the performance of the RF models.



	Overall Prevalence at *t0* [a]			RF Training Set [b]			RF Validation Set [c]		
Country	0 Pixels	1 Pixels	Prevalence of 1	n	Pre-existing BS pixels	Pre-existing Prevalence	N	Pre-existing BS pixels	Pre-existing Prevalence
Panama	8,901,004	23,805	0.27 %	27,482	43	0.16%	100,000	269	0.27%
Switzerland	6,816,510	193,939	2.77 %	100,000	1,464	0.2%	100,000	2,847	2.85%
Uganda	2,8231,555	26,831	0.10 %	67,600	30	1.5%	100,000	99	0.10%
Vietnam	40,108,425	146,260	0.36 %	100,000	199	0.04%	100,000	341	0.34%

a 0 indicates no detected presence of built-settlement and 1 indicates detected presence of built-settlement

b Pre-existing means that built-settlement was already in the sampled pixel, at time *t0,* selected for the random forest training set and therefore could never transition

c Pre-existing means that built-settlement was already in the pixel, at time *t0,* sampled for the random forest validation set and therefore could never transition

Section A8 – Modelling Times by Country

Given the 20 covariates detailed in the paper and predicting with four observed points the following were the computational times for the model as run on a local computer with 32GB RAM and an 8 core i7-6700 3.4GHz processor.

Table A6. Descriptions of the number of units (pixels) in given sample countries and the computational efficiency of the modelling runs.

Country	Total Pixels in Extent [a]	Total NonNA Pixels	Average Time per 100k non-NA pixels (secs)[b]	Total Time HH:MM:SS (secs)
Panama	20,700,555	8,924,809	71.6	01:46:27 (6,387)
Switzerland	13,056,459	7,010,449	13.9	01:05:30 (3,930)
Uganda	44,666,840	28,258,386	18.3	01:26:23 (5,183)
Vietnam	156,409,044	40,254,685	16.0	01:47:07 (6,427)

a Note: some of this pixels contain no data based upon the difference between the boundaries of the country and the rectangular extent of the raster

b The Panama times is much larger than other runs due to an unoptimized parameter that determines the number of blocks to divide the rasters into for the parallel prediction using the random forest. This parameter value has to be adjusted for a variety of considerations (total pixels, number of covariates, amount of no data values in raster extent, number of resulting blocks that can be skipped because they contain all no data values, etc.) with each model run. We did not optimize for Panama and Switzerland due to their small size and relative speed even when unadjusted.

Appendix B BSGMi versus BSGMα Considerations

The Global Built-Settlement Growth (GBSG) dataset [2] is a raster-based data product representing observed, interpolated, and projected global annual built-settlement (BS)[3] extents from 2000 through 2020 at 0.0008333 arc second resolution (~ 100m at the Equator). The GBSG was produced using an early version of the Built-Settlement Growth Model interpolation (BSGMi) and extrapolation (BSGMe) frameworks, which differ in a few significant ways from the BSGM framework validations presented in Nieves et al. (2020)[4] and Nieves et al. (2020). For the rest of the document, I refer to this early version of the BSGMi framework as the "BSGMi-α".

The BSGMi-α is largely similar to the BSGMi presented in Nieves et al. (2020), but with three key framework differences:

1. Only interpolating using information from two observed time points
2. The use of exponential growth/decay curves interpolating BS population counts and BS population density across time
3. When projecting forward, failure to account for edge cases which can result in uncontrolled BS growth in units with little to no BS at the last time of observation

Further, there are two model input parameter choices used in the generation of the GBSG that influence the resultant dataset and warrant keeping in mind when and before using the GBSG data products.

In this document, I cover the framework and input parameter differences, how they likely affect the datasets for end users, and recommended best practices in light of this. Here I do not cover the details of the entire BSGMi modelling framework and readers are referred to Nieves et al. (2020) and their corresponding supplementary materials for those details.

Framework Differences in Alpha Version

Interpolative Model Alpha (BSGMi- α): GBSG Data for 2000-2014

The BSGMi- α differs from the methods detailed in Nieves et al. (2020) in two key ways:

1. The use of exponential growth/decay curves to interpolate BS population and BS population density

2. The use of only two observed time points to interpolate BS population and BS population density

[2] https://www.worldpop.org/project/categories?id=15

[3] Built-settlement having the urban physical environment-based definition of "enclosed constructions above ground which are intended for or used 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" per (Pesaresi *et al.*, 2013, p. 2108)

[4] Preprint currently available under DOI: 10.20944/preprints201812.0250.v2

Appendix B

As described in Nieves et al. (2020), for every subnational unit *i*, the BS Pop is independently interpolated between every two observed BS time points, referred to as a “period,” with t_0 representing the first observed time point and t_1 representing the end observed time point of the period. This is where the procedure for the BSGMi- α diverges. First, the estimated BS population at the observed time points of the given period are transformed into an Urban Rural Ration (*URR*) using Equation 1:

$$URR_i = \frac{BSPOP_i}{POP_i - BSPOP_i} \qquad \text{[Eq. 1]}$$

with the BS population for unit *i* represented by $BSPOP_i$ and the total population for unit *i* represented by POP_i. For all time points *t* between t_0 and t_1, the URR values are interpolated using an exponential growth/decay formula, written in Equation 2 as:

$$URR_i(t) = URR_i(t_0) * e^{\overline{r_i} * t} \qquad \text{[Eq 2.]}$$

where the URR of the given unit *i* is given as a function of time *t* by $URR_i(t)$, $\overline{r_i}$ is the unit-average exponential rate of change across the number of evenly spaced time points between the two observed period end points, and *t* is the number of time points from the period initial time point t_0. $\overline{r_i}$ is calculated using Equation 3:

$$\overline{r_i} = \frac{\ln(URR(t_1)/URR(t_0))}{(t_1 - t_0)} \qquad \text{[Eq. 3]}$$

The interpolated URR values are then back-transformed into BS population counts by the relationship given in Equation 4:

$$BSPOP_i(t) = \; POP_i(t) * \frac{URR_i(t)}{1 + URR_i(t)} \qquad \text{[Eq. 4]}$$

prior to being used in the calculation of the time specific demand weights as detailed by Nieves et al. (2020).

Contrastingly, if provided more than two observed points in time, e.g. multiple periods, the procedure in Nieves et al. (2020) interpolates using a logistic growth/decay curve whose average unit-rate is determined using the information from all periods. Specifically, for each unit *i*, a logistic curve with a temporally dynamic carrying capacity is used, as shown in Equation 5:

$$BSPOP_i(t) = K_i(t) * \frac{e^{(\overline{r_i}+C_i)*t}}{1+e^{(\overline{r_i}+C_i)*t}} \quad \text{[Eq. 5]}$$

where $K_i(t)$ is the carrying capacity that varies with time, i.e. the time-specific unit total population, and the unit average logistic rate of change across all provided periods is represented by $\overline{r_i} + C_i$. $\overline{r_i} + C_i$ is estimated by fitting a linear regression across all observed BS population values for the given unit based upon the assumption, of the logistic growth curve with constraints, that the relationship in Equation 5 holds:

$$\{\ln\left(\frac{BSPOP_i(\{t_{OBS}\}}{K_i(\{t_{OBS}\})-BSPOP_i(\{t_{OBS}\})}\right)\} \cong \overline{r_i} + C_i \quad \text{[Eq. 6]}$$

where $\{t_{OBS}\}$ is the set of observed time points. Note how the left hand side of Eq. 6 is actually equal to a log transformed version of the URR in Eq. 1 given our choice of the total population to be the carrying capacity.

The first difference to note, is the manner in which the rate is calculated; with the BSGM-α using an exponential growth/decay function and only calculating the average rate based upon two observations, as opposed to the BSGMi method fitting a logistic growth/decay function over *two or more* observations. Additionally, the carrying capacity term in Eq. 5 and 6, allowing for the characteristic “s-shaped curve” of the logistic function, provides further differences in the estimates. With the logistic curve, if the ratio of BS population to the total population remains relatively small, then the resulting curve across time should be approximately exponential, i.e. concave upwards, in shape, but would undoubtedly have a different rate than if we were to fit across all observed points using the BSGM-α exponential approach. The fact that the carrying capacity term *K* in Eq 5 and 6 varies with time allows for potentially more complex outcomes of the interpolation, as seen in (Meyer and Ausubel, 1999b). Without delving too far into the numerous potential ways in which these terms can come to interact across both approaches, just looking at how a logistic curve with a static carrying capacity, we would expect the largest differences between the two methods for units that have relatively large proportions of BS population, i.e. highly urbanized areas, as the exponential (BSGM-α) approach assumes limitless growth, whereas the logistic approach assumes growth is constrained by some specified capacity. Whether the differences are meaningful in the output predicted extents, after being turned into weights and disaggregating the observed period changes across time, is highly locally dependent and would need to be investigated further.

Bibliography

Acuto, M., Parnell, S. and Seto, K. C. (2018) 'Building a global urban science', *Nature Sustainability*, 1(1), pp. 2–4. doi: 10.1038/s41893-017-0013-9.

Anas, A., Arnott, R. and Small, K. A. (1998) 'Urban Spatial Structure', *Journal of Economic Literature*, 36(3), pp. 1426–1464.

Angel, S. *et al.* (2011) 'The Dimensions of Global Urban Expansion: Estimates and Projections for All Countries, 2000-2050', *Progress in Planning*, 75, pp. 53–107. doi: 10.1016/j.progress.2011.04.001.

Angel, S., Sheppard, S. C. and Civco, D. L. (2005) *The Dynamics of Global Urban Expansion*. Washington, D. C.: The World Bank.

van Asselen, S. and Verburg, P. H. (2013) 'Land cover change or land-use intensification: simulating land system change with a global-scale land change model', *Global Change Biology*, 19(12), pp. 3648–3667. doi: 10.1111/gcb.12331.

Austin, A. L. and Brewer, J. W. (1971) 'World Population Growth and Related Technical Problems', *Technological Forecasting and Social Changes*, 3(1), pp. 23–49.

Balk, D. *et al.* (2004) *The Distribution of People and the Dimension of Place: Methodologies to Improve the Global Estimation of Urban Extents*. New York. Available at: https://pdfs.semanticscholar.org/1af9/f0c199753478311e2c6565e81ec75457eaf4.pdf (Accessed: 29 July 2018)

Balk, D. *et al.* (2006) 'Determining Global Population Distributions: Methods, Applications, and Data', *Advanced Parasitology*, 62, pp. 119–156.

Balk, D. *et al.* (2018) 'Understanding urbanization: A study of census and satellite-derived urban classes in the United States, 1990-2010', *PLOS ONE*. Edited by I. Benenson, 13(12), p. e0208487. doi: 10.1371/journal.pone.0208487.

Bartholomé, E. and Belward, A. S. (2005) 'GLC2000: a new approach to global land cover mapping from Earth observation data', *International Journal of Remote Sensing*, 26(9), pp. 1959–1977. doi: 10.1080/01431160412331291297.

Batty, M. (1997) 'Cellular Automata and Urban Form', *Journal of the American Planning Association*, 63(2), pp. 266–274.

Batty, M. (2008) 'Fifty years of urban modeling: Macro-statics to micro-dynamics', in Albeverio, S. et al. (eds) *The Dynamics of Complex Urban Systems: An Interdisciplinary Approach*. Heidelberg: Physica-Verlag, pp. 1–20.

Batty, M. (2009) 'Urban Modeling', in *International Encyclopedia of Human Geography*. Oxford, UK: Elsevier, pp. 51–58.

Batty, M. and Xie, Y. (1994) 'From Cells to Cities', *Environment and Planning B*, 21, pp. S31–S48.

Benenson, I. (2004) 'Agent-Based Modeling: From Individual Residential Choice to Urban Residential Dynamics', in Goodchild, M. F. and Janelle, D. G. (eds) *Spatially*

Integrated Social Science: Examples in Best Practice. Oxford University Press, pp. 67–95.

Benziger, V. (1996) ‘Urban Access and Rural Productivity Growth in Post-Mao China’, *Economic Development and Cultural Change*, 44(3), pp. 539–570. doi: 10.1086/452231.

Berechman, J. and Gordon, P. (1986) ‘Linked Models of Land-Use Transport Interactions: A Review’, in Hutchinson, B. and Batty, M. (eds) *Advances in Urban Systems Modelling*. Elsevier Ltd.

Bhaduri, B., Bright, E. and Coleman, P. (2007) ‘Landscan USA: a high resolution geospatial and temporal modeling approach for population distribution and dynamics’, *GeoJournal*, 69, pp. 103–177.

Bharti, N. *et al.* (2016) ‘Measuring populations to improve vaccination coverage’, *Scientific Reports*, 6(1), p. 34541. doi: 10.1038/srep34541.

de Boor, C. (2001) *A Practical Guide to Splines*. 2nd ed. New York: Springer-Verlag.

Booth, H. (2006) ‘Demographic forecasting: 1980 to 2005 in review’, *International Journal of Forecasting*, 22(3), pp. 547–581. doi: 10.1016/j.ijforecast.2006.04.001.

Box, G. E. P. and Jenkins, G. M. (1976) *Time Series Analysis: Forecasting and Control*. 2nd ed. San Francisco, CA, CA.

Bracken, I. and Martin, D. (1989) ‘The generation of spatial population distributions from census centroid data.’, *Environment and Planning A*, 21, pp. 537–543.

Breiman, L. (1996) ‘Bagging Predictors’, *Machine Learning*, 24(2), pp. 123–140.

Breiman, L. (2001a) ‘Random Forests’, *Machine Learning*, 45(1), pp. 5–32.

Breiman, L. (2001b) ‘Statistical Modeling: The Two Cultures’, *Statistical Science*, 16(3), pp. 199–231.

Brown de Colstoun, E. C. *et al.* (2017a) *Documentation for the Global Human Built-up And Settlement Extent (HBASE) Dataset From Landsat, v1*. Palisades, NY, NY.

Brown de Colstoun, E. C. *et al.* (2017b) *Documentation for the Global Man-made Impervious Surface (GMIS) Dataset From Landsat, v1*. Palisades, NY, NY.

Burgess, E. W. (1925) *The Growth of a City: An Introduction to a Research Project*. Chicago, IL, IL: University of Chicago Press. doi: 10.1080/00343404200021114.

Carr-Hill, R. (2013) ‘Missing Millions and Measuring Development Progress’, *World Development*, 46, pp. 30–44. doi: 10.1016/j.worlddev.2012.12.017.

Champion, T. and Hugo, G. (2017) *New Forms of Urbanization: Beyond the Urban-Rural Dichotomy*. 1st edn. Edited by T. Champion. Routledge. doi: 10.4324/9781315248073.

Chan, J. C. and Paelinckx, D. (2008) ‘Evaluation of Random Forest and Adaboost treebased ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery’, *Remote Sensing of Environment*, 112(6), pp. 2999–3011.

Chen, M., Mao, S. and Liu, Y. (2014) ‘Big Data: A Survey’, *Mobile Networks and Applications*, 19(2), pp. 171–209. doi: 10.1007/s11036-013-0489-0.

Cheriyadat, A. *et al.* (2007) ‘Mapping of settlements in high-resolution sattelite imagery using high performance computing’, *GeoJournal*, 69, pp. 119–129.

Clarke, K. C. and Gaydos, L. (1998) ‘Loose-coupling a Cellular Automaton Model and GIS: Long-term Urban Growth Prediction for San Francisco and Washington/Baltimore’, *International Journal of Geographic Information Sciences*, 12(7), pp. 699–714.

Clarke, K. C., Hoppen, S. and Gaydos, L. (1997) ‘A Self-modifying Cellular Automaton Model of Historical Urbanization in the San Francisco Bay Area’, *Environment and Planning B*, 24, pp. 247–261.

Cohen, B. (2004) ‘Urban growth in developing countries: A review of current trends and a caution regarding existing forecasting’, *World Development*, 32(1), pp. 23–51.

Cohen, B. (2006) ‘Urbanization in Developing Countries: Current Trends, Future Projections, and Key Challenges for Sustainability’, *Technology in Society*, 28, pp. 63–80.

Cohen, J. E. (1995) ‘Population Growth and Earth’s Human Carrying Capacity’, *Science*, 269(5222), pp. 341–346.

Corbane, C. *et al.* (2017) ‘Big earth data analytics on Sentinel-1 and Landsat imagery in support to global human settlements mapping’, *Big Earth Data*, 1(1–2), pp. 118–144. doi: 10.1080/20964471.2017.1397899.

Davis, K. (1965) ‘The Urbanization of the Human Population’, *Scientific American*, 213(3), pp. 40–53.

De Jong, P. and Tickle, L. (2006) ‘Extending Lee–Carter Mortality Forecasting’, *Mathematical Population Studies*, 13(1), pp. 1–18. doi: 10.1080/08898480500452109.

Deichmann, U., Balk, D. and Yetman, G. (2001) *Transforming Population Data for Interdisciplinary Usages: From census to grid*. Washington, D. C.

Diggle, P. J. (2014) *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns*. 3rd edn. CRC Press.

Dijkstra, L. and Poelman, H. (2014) *A harmonized definition of cities and rural areas: the new degree of urbanization*. WP 01/2014.

Douglass, M. (1989) ‘The Environmental Sustainability of Development: Coordination, Incentives and Political Will in Land Use Planning for the Jakarta Metropolis’, *Third World Planning Review*, 11(2), p. 211. doi: 10.3828/twpr.11.2.44113540kqt27180.

Doxsey-Whitfield, E. *et al.* (2015) 'Taking advantage of the improved availability of census data: A first look at the Gridded Population of the World, Version 4', *Papers in Applied Geography*, 1(3), pp. 226–234. doi: 10.1080/23754931.2015.1014272.

Dunn, O. J. (1964) 'Multiple comparisons using rank sums', *Technometrics*, 6, pp. 241–252.

Dyson, T. (2011) 'The role of the demographic transition in the process of urbanization', *Population and Development Review*, 37(Supplement), pp. 34–54.

Earth Observation Group, N. N. G. D. C. (2013) 'VIIRS Nighttime Lights - 2012 (Two Month Composite)'. Boulder, Colorado: NOAA National Centers for Environmental Information.

Ebenstein, A. and Zhao, Y. (2015) 'Tracking rural-to-urban migration in China: Lessons from the 2005 inter-census population survey', *Population Studies*, 69(3), pp. 337–353. doi: 10.1080/00324728.2015.1065342.

Eckert, S. and Kohler, S. (2014) 'Urbanization and Health in Developing Countries: A Systematic Review', *World Health & Population*, 15(1), pp. 7–20.

Ehrlich, D., Balk, D. and Sliuzas, R. (2020) 'Measuring and understanding global human settlements patterns and processes: innovation, progress and application', *International Journal of Digital Earth*, 13(1), pp. 2–8. doi: 10.1080/17538947.2019.1630072.

Eicher, C. L. and Brewer, C. A. (2001) 'Dasymetric mapping and areal interpolation: Implementation and evaluation', *Cartography and Geographic Information Science*, 28, pp. 125–138.

Epperson, J. F. (1987) 'On the Runge Example', *The American Mathematical Monthly*, 94(4), pp. 329–341.

ESA CCI (2017) 'European Space Agency Climate Change Initiative Landcover'. European Space Agency.

Esch, T *et al.* (2013) 'Urban Footprint Processor - Fully Automated Processing Chain Generating Settlement Masks from Global Data of the TanDEM-X Mission', *IEEE Geoscience and Remote Sensing Letters*, 10(6), pp. 1617–1621.

Esch, T. *et al.* (2018a) 'Where We Live—A Summary of the Achievements and Planned Evolution of the Global Urban Footprint', *Remote Sensing*, 10(6), p. 895. doi: 10.3390/rs10060895.

Esch, T. *et al.* (2018b) 'Exploiting big earth data from space – first experiences with the timescan processing chain', *Big Earth Data*, 2(1), pp. 36–55. doi: 10.1080/20964471.2018.1433790.

Esch, T. (2019) 'Personal Communication'.

European Commission and Statistical Office of the European Union (2019) *Methodological manual on territorial typologies: 2018 edition.* Available at: http://dx.publications.europa.eu/10.2785/930137 (Accessed: 17 August 2020).

European Space Agency (2013) 'Globcover, version 2.3'. ESA.

Ezeh, A. *et al.* (2017) ‘The history, geography, and sociology of slums and the health problems of people who live in slums’, *The Lancet*, 389(10068), pp. 547–558. doi: 10.1016/S0140-6736(16)31650-6.

Facebook Connectivity Lab and Columbia University Center for International Earth Science Information Network - CIESIN (2016) ‘High Resolution Settlement Layer’. New York: CIESIN.

Farrell, K. (2017) ‘The Rapid Urban Growth Triad: A New Conceptual Framework for Examining the Urban Transition in Developing Countries’, *Sustainability*, 9(8), p. 1407. doi: 10.3390/su9081407.

Farror, D. E. and Glauber, R. R. (1967) ‘Multicolinearity in regression analysis: The problem revisited’, *Review of Econometrics & Statistics*, 56(1), pp. 92–107.

Ferber, J. (1999) *Multi-agent Systems: An Introduction to Distributed Artificial Intelligence*. Harlow, UK, UK: Addison-Wesley.

Fildes, R. and Petropoulos, F. (2015) ‘Simple versus complex selection rules for forecasting many time series’, *Journal of Business Research*, 68, pp. 1692–1703.

Florczyk, A. J. *et al.* (2019) ‘The Generalised Settlement Area: mapping the Earth surface in the vicinity of built-up areas’, *International Journal of Digital Earth*, pp. 1–16. doi: 10.1080/17538947.2018.1550121.

Flowerdew, R., Green, M. and Evangelos, K. (1991) ‘Using areal interpolation methods in geographic information systems’, *Papers in Regional Science*, 70(3), pp. 303–315.

Forget, Y., Linard, C. and Gilbert, M. (2018) ‘Supervised Classification of Built-Up Areas in Sub-Saharan African Cities Using Landsat Imagery and OpenStreetMap’, *Remote Sensing*, 10(7), p. 1145. doi: 10.3390/rs10071145.

Freire, S. *et al.* (2015) ‘Combining GHSL and GPW to improve population mapping’, in *IEEE International Geoscience and Remote Sensing Symposium*, pp. 2541–2543. doi: 10.1109/IGARsS.2015.73263229.

Freire, S. *et al.* (2016) ‘Development of new open and free multi-temporal global population grids at 250m resolution’, in *19th AGILE Conference on Geographic and Information Science*. Available at: https://agile-online.org/conference_paper/cds/agile_2016/shortpapers/152_Paper_in_PDF.pdf.

Freire, S. *et al.* (2020) ‘Enhanced data and methods for improving open and free global population grids: putting “leaving no one behind” into practice’, *International Journal of Digital Earth*, 13(1), pp. 61–77. doi: 10.1080/17538947.2018.1548656.

Gao, J. and O’Neill, B. C. (2019) ‘Data-driven spatial modeling of global long-term urban land development: The SELECT model’, *Environmental Modelling & Software*, 119, pp. 458–471. doi: 10.1016/j.envsoft.2019.06.015.

Gao, J. and O’Neill, B. C. (2020) ‘Mapping global urban land for the 21st century with data-driven simulations and Shared Socioeconomic Pathways’, *Nature Communications*, 11(1), p. 2302. doi: 10.1038/s41467-020-15788-7.

Gaughan, Andrea E. *et al.* (2013) 'High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015', *PLoS One*, 8(2), p. e55882. doi: 10.1371/journal.pone.0055882.

Gaughan, Andrea E. *et al.* (2014) 'Exploring nationally and regionally defined models for large area population mapping', *International Journal of Digital Earth*. doi: 10.1080/17538947.2014.965761.

Gaughan, Andrea E. *et al.* (2016) 'Spatiotemporal patterns of population in mainland China, 1990 to 2010', *Scientific Data*, 3. doi: 10.1038/sdata.2016.5.

Gibson, J. and Li, C. (2017) 'THE ERRONEOUS USE OF CHINA'S POPULATION AND PER CAPITA DATA: A STRUCTURED REVIEW AND CRITICAL TEST', *Journal of Economic Surveys*, 31(4), pp. 905–922. doi: 10.1111/joes.12178.

Glaeser, E. L. (1998) 'Are Cities Dying?', *Journal of Economic Perspectives*, 12(2), pp. 139–160. doi: 10.1257/jep.12.2.139.

Goldewijk, K. K., Beusen, A. and Janssen, P. (2010) 'Long-term dynamic modeling of global population and built-up area in a spatially explicit way: HYDE 3.1', *The Holocene*, 20(4), pp. 565–573. doi: 10.1177/0959683609356587.

Gorelick, N. *et al.* (2017) 'Google earth engine: Planetary-scale geospatial analyss for everyone', *Remote Sensing of Environment*, 202, pp. 18–27.

Gottman, J. (1957) 'Megalopolis, or the urbanisation of the north eastern seaboard', *Economic Geography*, 33, pp. 189–200.

Green, N. (2007) 'Functional polycentricity: A formal definition in terms of social network analysis', *Urban Studies*, 44(11), pp. 2077–2103. doi: 10.1080/00420980701518941.

Haas, H. De (2010) 'Migration and Development: A Theoretical Perspective', *International Migration Review*, 44(1), pp. 227–264. doi: 10.1111/j.1747-7379.2009.00804.x.

Haklay, M. (2010) 'How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets', *Environment and Planning B: Urban Analytics and City Science*, 37(4), pp. 682–703. doi: 10.1068/b35097.

Hanjra, M. A. and Qureshi, M. E. (2010) 'Global Water Crisis and Future Food Security in an Era of Climate Change', *Food Policy*, 35, pp. 365–377. doi: 10.1016/j.foodpol.2010.05.006.

Harris, C. D. and Ullman, E. L. (1945) 'The nature of cities', *Annals of the American Academy of Political and Social Sciences*, 242, pp. 7–17.

Hay, S. I. *et al.* (2004) 'The Global Distribution and Population Risk of Malaria: Past, Present, and Future', *The Lancet Infectious Disease*, 4(6), pp. 327–226. doi: 10.1016/S1473-3099(04)01043-6.

Henderson, M. *et al.* (2003) 'Validation of urban boundaries derived from global night-time satellite imagery', *International Journal of Remote Sensing*, 24(3), pp. 595–609. doi: 10.1080/01431160304982.

Henderson, V. (2002) ‘Urbanization in Developing Countries’, *World Bank Research Observer*, 17(1), pp. 89–112.

Hijmans, R. J. *et al.* (2005) ‘Very high resolution interpolated climate surfaces for global land areas’, *International Journal of Climatology*, 25, pp. 1965–1978.

Hoalst-Pullen, N. and Patterson, M. W. (2011) ‘Applications and Trends of Remote Sensing in Professional Urban Planning’, *Geography Compass*, 5(5), pp. 249–261.

Hollander, J. B. and Németh, J. (2011) ‘The bounds of smart decline: a foundational theory for planning shrinking cities’, *Housing Policy Debate*, 21(3), pp. 349–367. doi: 10.1080/10511482.2011.585164.

Holm, S. (1979) ‘A simple sequentially rejective multiple test procedure’, *Scandanavian Journal of Statistics*, 6(2), pp. 65–70.

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

Hyndman, R. J. *et al.* (2002) ‘A state space framework for automatic forecasting using exponential smoothing methods’, *International Journal of Forecasting*, 18(3), pp. 439–454.

Hyndman, R. J. and Booth, H. (2008) ‘Stochastic population forecasts using functional data models for mortality, fertility and migration’, *International Journal of Forecasting*, 24(3), pp. 323–342. doi: 10.1016/j.ijforecast.2008.02.009.

Hyndman, R. J. and Khandakar, Y. (2008) ‘Automatic Time Series Forecasting: The forecast package for R’, *Journal of Statistical Software*, 27(3).

Hyndman, R. J. and Shahid Ullah, Md. (2007) ‘Robust forecasting of mortality and fertility rates: A functional data approach’, *Computational Statistics & Data Analysis*, 51(10), pp. 4942–4956. doi: 10.1016/j.csda.2006.07.028.

Illian, J. B., Sørbye, S. H. and Rue, H. (2012) ‘A toolbox for fitting complex spatial point process models using integrated nested Laplace approximation (INLA)’, *The Annals of Applied Statistics*, 6(4), pp. 1499–1530. doi: 10.1214/11-AOAS530.

Institute for Health Metrics and Evaluation (2020) *COVID-19 model FAQs*, *healthdata.org*. Available at: http://www.healthdata.org/covid/faqs (Accessed: 19 June 2020).

Jilge, M. *et al.* (2019) ‘Gradients in urban material composition: A new concept to map cities with spaceborne imaging spectroscopy data’, *Remote Sensing of Environment*, 223, pp. 179–193. doi: 10.1016/j.rse.2019.01.007.

Jones, B., Balk, D. and Leyk, S. (2020) ‘Urban Change in the United States, 1990–2010: A Spatial Assessment of Administrative Reclassification’, *Sustainability*, 12(4), p. 1649. doi: 10.3390/su12041649.

Joss, S., Cowley, R. and Tomozeiu, D. (2013) ‘Towards the “ubiquitous eco-city”: An analysis of the internationalisation of eco-city policy and practice’, *Urban Research & Practice*, 6(1), pp. 54–74. doi: 10.1080/17535069.2012.762216.

Juran, S. *et al.* (2018) ‘Geospatial mapping of access to timely essential surgery in sub-Saharan Africa’, *BMJ Global Health*, 3(4), p. e000875. doi: 10.1136/bmjgh-2018-000875.

Kamusoko, C. and Gamba, J. (2015) ‘Simulating Urban Growth Using a Random Forest-Cellular Automata (RF-CA) Model’, *ISPRS International Journal of Geo-Information*, 4, pp. 447–470.

Kelly, P. F. (1998) ‘The politics of urbanrural relations: land use conversion in the Philippines’, *Environment and Urbanization*, 10(1), pp. 35–54. doi: 10.1177/095624789801000116.

Kloosterman, R. C. and Musterd, S. (2001) ‘The polycentric urban region: Towards a research agenda’, *Urban Studies*, 38(4), pp. 623–633.

Koning, G. H. J. de *et al.* (1999) ‘Multi-scale modelling of land use change dynamics in Ecuador’, *Agricultural Systems*, 61, pp. 77–93.

Korale, R. B. M. (2002) *Post Enumeration Survey 2001 [Nepal Population Census] Draft Report*. Kathmandu, Nepal.

Kruskal, W. H. and Wallis, W. A. (1952) ‘Use of ranks in one-criterion variance analysis’, *Journal of the American Statistical Association*, 47, pp. 583–621.

Kuffer, M., Barros, J. and Sliuzas, R. V. (2014) ‘The development of a morphological unplanned settlement index using very-high-resolution (VHR) imagery’, *Computers, Environment and Urban Systems*, 48, pp. 138–152. doi: 10.1016/j.compenvurbsys.2014.07.012.

Kuffer, M., Pfeffer, K. and Sliuzas, R. (2016) ‘Slums from Space—15 Years of Slum Mapping Using Remote Sensing’, *Remote Sensing*, 8(6), p. 455. doi: 10.3390/rs8060455.

Lamarche, C. *et al.* (2017) ‘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 Sensing*, 9(36). doi: 10.3390/rs9010036.

Langford, M., Maguire, D. J. and Unwin, D. J. (1991) ‘The areal interpolation problem: Estimating population using remote sensing in a GIS framework’, in *Handling Geographic Information*. Essex, U.K.: Longman Scientific and Technical, pp. 55–77.

Leao, S., Bishop, I. and Evans, D. (2004) ‘Simulating Urban Growth in a Developing Nation’s Region Using a Cellular Automata-based Model’, *Journal of Urban Planning and Development*, 130(3), pp. 145–158.

Ledent, J. (1982) ‘Rural-Urban Migration, Urbanization, and Economic Development’, *Economic Development and Cultural Change*, 30(3), pp. 507–538.

Lehner, B., Verdin, K. and Jarvis, A. (2008) ‘New Global Hydrography Derived from Spaceborne Elevation Data’, *Eos, Transactions of the American Geophysical Union*, 89(10), pp. 93–94. doi: 10.1029/2008EO100001.

Leyk, S. *et al.* (2019) ‘The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use’, *Earth System Science Data*, 11(3), pp. 1385–1409. doi: 10.5194/essd-11-1385-2019.

Li, X. and Gong, P. (2016) ‘Urban growth models: progress and perspective’, *Science Bulletin*, 61(21), pp. 1637–1650. doi: 10.1007/s11434-016-1111-1.

Li, Y. *et al.* (2016) ‘Forecasting the daily power output of a grid-connected photovoltaic system based on multivariate adaptive regression splines’, *Applied Energy*, 180, pp. 392–401. doi: 10.1016/j.apenergy.2016.07.052.

Liaw, A. and Wiener, M. (2002) ‘Classification and Regression by randomForest’, *R News*, 3(2), pp. 18–22.

Linard, C., Alegana, Victor A., *et al.* (2010) ‘A high resolution spatial population database of Somalia for disease risk mapping’, *International Journal of Health Geographics*, 9(1), p. 45. doi: 10.1186/1476-072X-9-45.

Linard, C., Alegana, Victor A, *et al.* (2010) ‘A high resolution spatial population database of Somalia for disease risk mapping’, *International Journal of Health Geographics*, 9(1), p. 45. doi: 10.1186/1476-072X-9-45.

Linard, C. *et al.* (2012) ‘Population Distribution, Settlement Patterns and Accessibility across Africa in 2010’, *PLoS ONE*. Edited by G. J.-P. Schumann, 7(2), p. e31743. doi: 10.1371/journal.pone.0031743.

Linard, C. *et al.* (2014) ‘Use of active and passive VGI data for population distribution modelling: experience from the WorldPop project’, in *Proc. of the Eighth International Conference on Geographic Information Science*. Vienna, Austria, pp. 1–16.

Linard, C. *et al.* (2017) ‘Modelling changing population distributions: an example of the Kenyan Coast, 1979–2009’, *International Journal of Digital Earth*, 10(10), pp. 1017–1029. doi: 10.1080/17538947.2016.1275829.

Linard, C., Gilbert, M. and Tatem, A. J. (2013) ‘Assessing the use of global land cover data for guiding large area population distribution modelling’, *GeoJournal*, 76(5), pp. 525–538. doi: 10.1007/s10708-010-9364-8.

Linard, C., Tatem, A. J. and Gilbert, M. (2013) ‘Modelling Spatial Patterns of Urban Growth in Africa’, *Applied Geography*, 44, pp. 23–32.

Liu, Y. and Feng, Y. (2012) ‘A Logistic Based Cellular Automata Model for Continuous Urban Growth Simulation: A Case Study of the Gold Coast City, Australia’, in Heppenstall, A. J. (ed.) *Agent-Based Models of Geographical Systems*. Springer, pp. 643–662.

Lloyd, C. T. *et al.* (2019) ‘Global spatio-temporally harmonised datasets for producing high-resolution gridded population distribution datasets’, *Big Earth Data*, 3(2), pp. 108–139. doi: 10.1080/20964471.2019.1625151.

Lucci, P., Bhatkal, T. and Khan, A. (2018) ‘Are we underestimating urban poverty?’, *World Development*, 103, pp. 297–310. doi: 10.1016/j.worlddev.2017.10.022.

Martin, D. (1989) ‘Mapping population data from zone centroid locations’, *Transactions of the Institute of British Geographers*, 14, pp. 90–97.

Martin, D. and Bracken, I. (1991) ‘Techniques for modelling population-related raster datasets’, *Environment and Planning A*, 23, pp. 1069–1075.

McCullagh, P. and Nelder, J. A. (1989) 'The components of a generalized linear model', in *Generalized Linear Models*. Chapan & Hall/CRC.

McDonald, R. I. *et al.* (2011) 'Urban growth, climate change, and freshwater availability', *Proceedings of the National Academy of Sciences*, 108(15), pp. 6312–6317. doi: 10.1073/pnas.1011615108.

McGranahan, G., Balk, D. and Anderson, B. (2007) 'The Rising Tide: Assessing the Risks of Climate Change and Human Settlements in Low Elevation Coastal Zones', *Environment & Urbanization*, 19(1), pp. 17–37. doi: 10.1177/0956247807076960.

McKee, J. J. *et al.* (2015) 'Locally adaptive, spatially explicit projection of US population for 2030 and 2050', *Proceedings of the National Academy of Sciences*, 112(5), pp. 1344–1349. doi: 10.1073/pnas.1405713112.

McNeil, D. R., Trussell, T. J. and Turner, J. C. (1977) 'Spline Interpolation of Demographic Data', *Demography*, 14(2), pp. 245–252.

Mennis, J. (2003) 'Generating surface models of population using dasymetric mapping', *Professional Geographer*, 55(1), pp. 31–42.

Mennis, J. and Hultgren, T. (2006) 'Intelligent dasymetric mapping and its application to areal interpolation', *Cartography and Geographic Information Science2*, 33, pp. 179–194.

Merriam-Webster (2019) *urban*, *Merriam-Webster Dictionary*. Available at: https://www.merriam-webster.com/dictionary/urban (Accessed: 7 February 2019).

Meyer, P. S. and Ausubel, J. H. (1999a) 'Carrying capacity: A model with logistically varying limits', *Technological Forecasting and Social Change*, 61(3), pp. 209–214. doi: 10.1016/S0040-1625(99)00022-0.

Meyer, P. S. and Ausubel, J. H. (1999b) 'Carrying capacity: A model with logistically varying limits', *Technological Forecasting and Social Change*, 61(3), pp. 209–214. doi: 10.1016/S0040-1625(99)00022-0.

Meyer, W. B. and Turner, B. L. (1992) 'Human Population Growth and Global Land-Use / Cover Change', *Annual Review of Ecology and Systematics*, 23(1992), pp. 39–61. doi: 10.2307/2097281.

Microsoft (2018) *US Building Footprints*, *GitHub*. Available at: https://github.com/Microsoft/USBuildingFootprints (Accessed: 7 February 2019).

Montgomery, M. R. *et al.* (2003) *Cities Transformed: Demographic Change and Its Implications in the Developing World*. Washington, D.C.: National Academies Press, p. 10693. doi: 10.17226/10693.

Nagle, N. N. *et al.* (2014) 'Dasymetric Modeling and Uncertainty', *Annals of the Association of American Geographers*, 104(1), pp. 80–95. doi: 10.1080/00045608.2013.843439.

Neis, P. and Zipf, A. (2012) 'Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap', *ISPRS International Journal of Geo-Information*, 1, pp. 146–165. doi: 10.3390/ijgi1020146.

Nelder, J. A. and Wedderburn, R. W. M. (1972) ‘Generalized Linear Models’, *Journal of the Royal Statistical Society Series A*, 135(Part 3), pp. 370–384.

Nieves, Jeremiah J. *et al.* (2017) ‘Examining the correlates and drivers of human population distributions across low- and middle-income countries’, *Journal of The Royal Society Interface*, 14(137), p. 20170401. doi: 10.1098/rsif.2017.0401.

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

Nieves, Jeremiah J. *et al.* (2020b) ‘Predicting Near-Future Built-Settlement Expansion Using Relative Changes in Small Area Populations’, *Remote Sensing*, 12(10), p. 1545. doi: 10.3390/rs12101545.

Oliver, F. R. (1964) ‘Methods of Estimating the Logistic Growth Function’, *Royal Statistical Society C*, 13(2), pp. 57–66.

Openshaw, S. (1984) ‘The modifiable areal unit problem’, *Concepts and Techniques in Modern Geography*, 38.

OpenStreetMap Contributers (2017) ‘OpenStreetMap (OSM) Database’. OSM.

Ord, J. K., Koehler, A. B. and Snyder, R. D. (1997) ‘Estimation and Prediction for a Class of Dynamic Nonlinear Statistical Models’, *Journal of the American Statistical Association*, 92(1621–1629).

Oswalt, P., Rieniets, T. and Schirmel, H. (2006) ‘Atlas of shrinking cities’. Ostfildern: Hatje Cantz.

Parr, J. (2004) ‘The polycentric urban region: A closer inspection’, *Regional Studies*, 38(3), pp. 231–240.

Patel, N. *et al.* (2015) ‘Multitemporal Settlement and Population Mapping From Landsat Using Google Earth Engine’, *International Journal of Applied Earth Observation and Geoinformation*, 35(Part B), pp. 199–208. doi: 10.1016/j.jag.2014.09.005.

Pegels, C. C. (1969) ‘Exponential Forecasting: Some New Variations’, *Management Science*, 15(5), pp. 311–315.

Pesaresi, M. *et al.* (2013) ‘A Global Human Settlement Layer from Optical HR/VHR Remote Sensing Data: Concept and First Results’, *IEEE Journal of Selected Topics in Applied Earth Observation & Remote Sensing*, 6(5), pp. 2102–2131. doi: 10.1109/JSTARS.2013.2271445.

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*. Publications Office of the European Union.

Pezzulo, C. *et al.* (2017) ‘Sub-national mapping of population pyramids and dependency ratios in Africa and Asia’, *Scientific Data*, 4, p. 170089. doi: 10.1038/sdata.2017.89.

van Poppel, F. and van der Heijden, C. (1997) ‘The effects of water supply on infant and childhood mortality: a review of historical evidence’, *Health Transition*

Review: The Cultural, Social, and Behavioural Determinants of Health, 7(2), pp. 113–148.

Potere, D. *et al.* (2009) ‘Mapping urban areas on a global scale: which of the eight maps now available is more accurate?’, *International Journal of Remote Sensing*, 30(24), pp. 6531–6558. doi: 10.1080/01431160903121134.

Potere, D. and Schneider, A. (2007) ‘A critical look at representations of urban areas in global maps’, *GeoJournal*, 69(1–2), pp. 55–80. doi: 10.1007/s10708-007-9102-z.

Pozzi, F. and Small, C. (2005) ‘Analysis of urban land cover and population density in the United States’, *Photogrammetric Engineering & Remote Sensing2*, 71, pp. 719–726.

Preston, S. H. and van de Walle, E. (1978) ‘Urban French Mortality in the Nineteenth Century’, *Population Studies*, 32(2), p. 275. doi: 10.2307/2173562.

R Core Team (2017) ‘R: A Language and Environment Layer for Statistical Computing’. Vienna, Austria: R Foundation for Statistical Computing.

R Core Team (2019) ‘R: A Language and Environment Layer for Statistical Computing’. Vienna, Austria: R Foundation for Statistical Computing.

Reed, F. *et al.* (2018) ‘Gridded Population Maps Informed by Different Built Settlement Products’, *Data*, 3(3), p. 33. doi: 10.3390/data3030033.

Rodriguez-Galiano, V. F. *et al.* (2012) ‘An assessment of the effectiveness of a random forest classifier for landcover detection’, *Photogrammetry & Remote Sensing2*, 67, pp. 93–104.

Rogan, W. J. and Gladen, B. (1978) ‘Estimating prevalence from the results of a screening test’, *American Journal of Epidemiology*, 107(1), pp. 71–76.

Rosner, B. (2011) ‘Multisample Inference’, in Taylor, M. (ed.) *Fundamentals of Biostatistics*. 7th edn. Boston, MA: Brooks/Cole, pp. 516–576.

Runge, C. (1901) ‘Uber empirische Funktionen und die Interpolation zwischen aquidistanten Ordinaten’, *Zeitschrift fur Mathematik und Physik*, 46, pp. 227–243.

Sabry, S. (2010) ‘How poverty is underestimated in Greater Cairo, Egypt’, *Environment and Urbanization*, 22(2), pp. 523–541. doi: 10.1177/0956247810379823.

Sakoda, J. M. (1971) ‘The Checkerboard Model of Social Interaction’, *Journal of Mathematical Sociology*, 1, pp. 119–132.

Sante, I. *et al.* (2010) ‘Cellular Automata Models for the Simulation of Real-world Urban Processes: A Review and Analysis’, *Landscape and Urban Planning*, 96, pp. 108–122. doi: 10.1016/j.landurbplan.2010.03.001.

Schaldach, R. *et al.* (2011) ‘An integrated approach to modelling land-use change on continental and global scales’, *Environmental Modelling & Software*, 26(8), pp. 1041–1051. doi: 10.1016/j.envsoft.2011.02.013.

Schapire, R. E. (2013) 'Explaining Adaboost', in Schölkopf, B., Luo, Z., and Vovk, V. (eds) *Empirical Inference*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-41136-6.

Schelling, T. (1971) 'Dynamic Models of Segregation', *Journal of Mathematical Sociology*, 1, pp. 143–186.

Schneider, A. *et al.* (2003) 'Mapping Urban Areas by Fusing Multiple Sources of Coarse Resolution Remotely Sensed Data', *Photogrammetry & Remote Sensing*, 69(12), pp. 1377–1386.

Schneider, A., Friedl, M. A. and Potere, D. (2010) 'Mapping Urban Areas Using MODIS 500-m Data: New Methods and Datasets Based on "Urban Ecoregions"', *Remote Sensing of the Environment*, 114, pp. 1733–1746.

Schneider, A and Woodcock, C. E. (2008) 'Compact, Dispersed, Fragmented, Extensive? A Comparison of Urban Growth in Twenty-five Global Cities using Remotely Sensed Data, Pattern Metrics and Census Information', *Urban Studies*, 45(3), pp. 659–692. doi: 10.1177/0042098007087340.

Schneider, A. and Woodcock, C. E. (2008) 'Compact, Dispersed, Fragmented, Extensive? A Comparison of Urban Growth in Twenty-five Global Cities using Remotely Sensed Data, Pattern Metrics and Census Information', *Urban Studies*, 45(3), pp. 659–692. doi: 10.1177/0042098007087340.

Schroeder, J. P. (2007) 'Target-density Weighting Interpolation and Uncertainty Evaluation for Temporal Analysis of Census Data', *Geographical Analysis*, 39, pp. 311–335.

Scott, G. and Rajabifard, A. (2017) 'Sustainable Development and Geospatial Information: A Strategic Framework for Integrating a Global Policy Agenda into National Geospatial Capabilities', *Geo-spatial Information Science*, 20(2), pp. 59–76.

Seto, K. C. *et al.* (2011) 'A Meta-Analysis of Global Urban Land Expansion', *PLoS One*, 6(8), p. e23777. doi: 10.1371/journal.pone.0023777.

Seto, K C, Guneralp, B. and Hutyra, L. R. (2012) 'Global Forecasts of Urban Expansion to 2030 and Direct Impacts on Biodiversity and Carbon Pools', *Proceedings of the National Academy of Sciences of the United States of America*, 109(40), pp. 16083–16088. doi: 10.1073/pnas.1211658109.

Seto, K. C., Guneralp, B. and Hutyra, L. R. (2012) 'Global Forecasts of Urban Expansion to 2030 and Direct Impacts on Biodiversity and Carbon Pools', *Proceedings of the National Academy of Sciences of the United States of America*, 109(40), pp. 16083–16088. doi: 10.1073/pnas.1211658109.

Shmueli, G. (2010) 'To Explain or Predict', *Statistical Science*, 25(3), pp. 289–310. doi: 10.1214/10-STS330.

Sibly, R. M., Barker, D., Denham, M. C., Hone, J. and Pagel, M (2005) 'On the Regulation of Populations of Mammals, Birds, Fish, and Insects', *Science*, 309(5734), pp. 607–610. doi: 10.1126/science.1110760.

Sibly, R. M., Barker, D., Denham, M. C., Hone, J. and Pagel, M. (2005) 'On the Regulation of Populations of Mammals, Birds, Fish, and Insects', *Science*, 309(5734), pp. 607–610. doi: 10.1126/science.1110760.

Sinha, P. *et al.* (2019) ‘Assessing the spatial sensitivity of a random forest model: Application in gridded population modeling’, *Computers, Environment and Urban Systems*, 75, pp. 132–145. doi: 10.1016/j.compenvurbsys.2019.01.006.

Small, C. (2009) ‘The color of cities:An overview of urban spectral diversity’, in Herold, M. and Gamba, P. (eds) *Global Mapping of Human Settlements*. New York: Taylor & Francis, pp. 59–106.

Small, C. *et al.* (2011) ‘Spatial scaling of stable night lights’, *Remote Sensing of Environment*, 115(2), pp. 269–280. doi: 10.1016/j.rse.2010.08.021.

Small, C. (2016) ‘Projecting the Urban Future: Contributions from Remote Sensing’, *Spatial Demography*, 4(1), pp. 17–37. doi: 10.1007/s40980-015-0002-4.

Small, C. *et al.* (2018) ‘Decades of urban growth and development on the Asian megadeltas’, *Global and Planetary Change*, 165, pp. 62–89. doi: 10.1016/j.gloplacha.2018.03.005.

Small, C. and Cohen, J. E. (2004) ‘Continental physiography, climate, and the global distribution of human population’, *Current Anthropology*, 45, pp. 269–277.

Small, C. and Nicholls, R. J. (2003) ‘A global analysis of human settlement in coastal zones’, *Coastal Research*, 19(3), pp. 584–599.

Small, C., Pozzi, F. and Elvidge, C. D. (2005) ‘Spatial analysis of global urban extent from DMSP-OLS night lights’, *Remote Sensing of Environment*, 96, pp. 277–291. doi: 10.1016/j.rse.2005.02.002.

Smith, S. K. (1997) ‘Further thoughts on simplicity and complexity in population projection models’, *International Journal of Forecasting*, 13, pp. 557–565.

Solecki, W., Seto, K. C. and Marcotullio, P. J. (2013) ‘It’s Time for an Urbanization Science’, *Environment: Science and Policy for Sustainable Development*, 55(1), pp. 12–17. doi: 10.1080/00139157.2013.748387.

Sorichetta, Alessandro *et al.* (2015) ‘High-resolution gridded population distribution datasets of Latin America in 2010, 2015, and 2020’, *Scientific Data*, 2, p. 150045. doi: 10.1038/sdata.2015.45.

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

Steele, J. E. *et al.* (2017) ‘Mapping poverty using mobile phone and satellite data’, *Journal of The Royal Society Interface*, 14(127), p. 20160690. doi: 10.1098/rsif.2016.0690.

Stephenson, J., Newman, K. and Mayhew, S. (2010) ‘Population dynamics and climate change: What are the links?’, *Journal of Public Health*, 32(2), pp. 150–156. doi: 10.1093/pubmed/fdq038.

Stevens, F R *et al.* (2015) ‘Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-sensed Data and Ancillary Data’, *PLoS One*, 10(2), p. e0107042. doi: 10.1371/journal.pone.0107042.

Stevens, F. R. *et al.* (2020) 'Comparisons of two global built area land cover datasets in methods to disaggregate human population in eleven countries from the global South', *International Journal of Digital Earth*, 13(1), pp. 78–100. doi: 10.1080/17538947.2019.1633424.

Strobl, C. *et al.* (2007) 'Bias in random forest variable importance measures: Illustrations, sources and a solution', *BMC Bioinformatics*, 8(1), p. 25. doi: 10.1186/1471-2105-8-25.

Strobl, C. *et al.* (2008) 'Conditional variable importance for random forests', *BMC Bioinformatics*, 9(1), p. 307. doi: 10.1186/1471-2105-9-307.

Sverdlik, A. (2011) 'Ill-health and poverty: A literature review on health in informal settlements', *Environment and Urbanization*, 23(1), pp. 123–155. doi: 10.1177/0956247811398604.

Tatem, A. J. *et al.* (2007) 'High Resolution Population Maps for Low Income Nations: Combining Land Cover and Census in East Africa', *PLoS ONE*, 2, p. e1298.

Tatem, A. J. (2014) 'Mapping the denominator:Spatial demography in the measurement of progress', *International Health*, 6(3), pp. 153–155. doi: 10.1093/inthealth/ihu057.

Tatem, A. J. (2018) 'Innovation to impact in spatial epidemiology', *BMC Medicine*, 16(1), p. 209. doi: 10.1186/s12916-018-1205-5.

Tatem, A. and Linard, C. (2011) 'Population mapping of poor countries', *Nature*, 474(7349), pp. 36–36. doi: 10.1038/474036d.

Tayyebi, A. *et al.* (2013) 'Hierarchical modeling of urban growth across the conterminous USA: Developing meso-scale quantity drivers for the Land Transformation Model', *Journal of Land Use Science*, 8(4), pp. 422–442. doi: 10.1080/1747423X.2012.675364.

The World Bank (2020) *The World By Income and Region*, *worldbank.org*. Available at: https://datatopics.worldbank.org/world-development-indicators/the-world-by-income-and-region.html (Accessed: 14 December 2019).

Von Thunen, J. H. (1966) *Von Thunen's 'Isolated State': An English Translation of 'Der Isolierte Staat'*. Edited by C. M. Wartenberg and P. Hall. Oxford, UK, UK: Pergamon Press.

Tobler, W. (1970) 'A Computer Movie Simulating Urban Growth in the Detroit Region', *Economic Geography*, 46(Supplement: Proceedings of the International Geographical Union Commission on Quantitative Methods), pp. 234–240. doi: 10.1126/science.11.277.620.

UCL Geomatics (2017) *Land Cover CCI Product User Guide Version 2.0*.

Ugarte, M. *et al.* (2012) 'Projections of cancer mortality risks using spatio-temporal P-spline models', *Statistical Methods in Medical Research*, 21(5), pp. 545–560. doi: 10.1177/0962280212446366.

U.N. Enviroment Programme World Conservation Monitoring Centre and IUCN World Commission on Protected Areas (2015) 'World Database on Protected Areas'. IUCN & UNEP.

United Nations (2015) *World Urbanization Prospects: The 2014 Revision*. ST/ESA/SER.A/366. New York, New York: United Nations, Dept. of Economic and Social Affairs, Population Division.

United Nations (2016) *Transforming Our World: The 2030 Agenda for Sustainable Development*.

United Nations (2018) *World Urbanization Prospects: The 2018 Revision*. New York.

United Nations (2019) 'World Population Prospects 2019: Highlights'. U.N, Department of Economic and Social Affairs, Population Division.

United Nations - Economic and Social Council (2016) *Report of the high-level political forum on sustainable development convened under the auspices of the Economic and Social Council at its 2016 session*.

Verburg, P. H. *et al.* (1999) 'A spatial explicit allocation procedure for modelling the pattern of land use change based upon actual land use', *Ecological Modelling*, 116, pp. 45–61.

Verburg, P. H. *et al.* (2002) 'Modeling the Spatial Dynamics of Regional Land Use: The CLUE-S Model', *Environmental Management*, 30(3), pp. 391–405. doi: 10.1007/s00267-002-2630-x.

Verburg, P. H. *et al.* (2004) 'Landuse Change Modelling: Current Practice and Research Priorities', *GeoJournal*, 61, pp. 309–324.

Verburg, P. H. and Overmars, K. P. (2009) 'Combining top-down and bottom-up dynamics in land use modeling: exploring the future of abandoned farmlands in Europe with the Dyna-CLUE model', *Landscape Ecology*, 24(9), pp. 1167–1181. doi: 10.1007/s10980-009-9355-7.

Verhulst, P.-F. (1838) 'Notice sur la loi que la population poursuit dans son accroissement', *Correspondance Mathe-matique et Physique*, 10, pp. 113–121.

Vlahov, D. *et al.* (2007) 'Urban as a Determinant of Health', *Journal of Urban Health*, 84(S1), pp. 16–26. doi: 10.1007/s11524-007-9169-3.

van Vliet, J., Eitelberg, D. A. and Verburg, P. H. (2017) 'A global analysis of land take in cropland areas and production displacement from urbanization', *Global Environmental Change*, 43, pp. 107–115. doi: 10.1016/j.gloenvcha.2017.02.001.

Vogelmann, J. E. *et al.* (2001) 'Completion of the 1990s National Land Cover Data Set for the Conterminous United States from Landsat Thematic Mapper Data and Ancillary Data Sources', *Photogrammetric Engineering & Remote Sensing*, 67, pp. 650–662.

Wardrop, N. A. *et al.* (2018) 'Spatially disaggregated population estimates in the absence of national population and housing census data', *Proceedings of the National Academy of Sciences*, 115(14), pp. 3529–3537. doi: 10.1073/pnas.1715305115.

Weber, E. M. *et al.* (2018) 'Census-independent population mapping in northern Nigeria', *Remote Sensing of Environment*, 204, pp. 786–798. doi: 10.1016/j.rse.2017.09.024.

White, R. and Engelen, G. (1997) ‘Cellular Automata as the Basis of Integrated Dynamic Regional Modelling’, *Environment and Planning B*, 24, pp. 235–246.

White, R. and Engelen, G. (2000) ‘High Resolution Modelling of the Spatial Dynamics of Urban and Regional Systems’, *Computers, Environment, and Urban Systems*, 24(383–400).

Wickham, H. (2014) ‘Tidy Data’, *Journal of Statistical Software*, 59(10). doi: 10.18637/jss.v059.i10.

Wilcoxon, F. (1945) ‘Individual Comparisons By Ranking Methods’, *Biometrics*, 1, pp. 80–83.

Wilson, A. G. (1976) ‘Catastrophe Theory and Urban Modelling: An Application to Modal Choice’, *Environment and Planning A*, 8(3), pp. 351–356. doi: https://doi.org/10.1068/a080351.

Wolfram, S. (1984) ‘Cellular automata as models of complexity’, *Nature*, 311, pp. 419–424.

WorldPop - School of Geography and Environmental Science - University of Southampton; *et al.* (2018) ‘Global High Resolution Population Denominators Project’. Bill and Melinda Gates Foundation (OPP1134076). doi: https://dx.doi.org/10.5258/SOTON/WP00645.

Wright, J. K. (1936) ‘A method of mapping densities of population’, *The Geographical Review*, 26, pp. 103-110#.

Yang, L. *et al.* (2003) ‘Urban Land-Cover Change Detection through Sub-Pixel Imperviousness Mapping Using Remotely Sensed Data’, 69(9), pp. 1003–1010.

Zelinsky, W. (1971) ‘The Hypothesis of the Mobility Transition’, *Geographical Review*, 61(2), pp. 219–249.

Zhang, Q., Pandey, B. and Seto, K. C. (2016) ‘A Robust Method to Generate a Consistent Time Series From DMSP/OLS Nighttime Light Data’, *IEEE Transactions on Geoscience and Remote Sensing*, 54(10), pp. 5821–5831. doi: 10.1109/TGRS.2016.2572724.

Zhu, Z. *et al.* (2019) ‘Understanding an urbanizing planet: Strategic directions for remote sensing’, *Remote Sensing of Environment*, 228, pp. 164–182. doi: 10.1016/j.rse.2019.04.020.

Zoraghein, H. and Leyk, S. (2019) ‘Data-enriched interpolation for temporally consistent population compositions’, *GIScience & Remote Sensing*, 56(3), pp. 430–461. doi: 10.1080/15481603.2018.1509463.