International Journal of Digital Earth

ISSN: 1753-8947 (Print) 1753-8955 (Online) Journal homepage: https://www.tandfonline.com/loi/tjde20

Detecting and mapping slums using open data: a case study in Kenya

Ron Mahabir, Peggy Agouris, Anthony Stefanidis, Arie Croitoru & Andrew T. Crooks

To cite this article: Ron Mahabir, Peggy Agouris, Anthony Stefanidis, Arie Croitoru & Andrew T. Crooks (2018): Detecting and mapping slums using open data: a case study in Kenya, International Journal of Digital Earth, DOI: 10.1080/17538947.2018.1554010

To link to this article: https://doi.org/10.1080/17538947.2018.1554010

Published online: 04 Dec 2018.

INTERNATIONAL JOURNAL OF DIGITAL EARTH
https://doi.org/10.1080/17538947.2018.1554010

Detecting and mapping slums using open data: a case study in Kenya

Ron Mahabir [a,b], Peggy Agouris[a,c], Anthony Stefanidis [a,b,d], Arie Croitoru [a,b] and Andrew T. Crooks [a,b,e]

[a]Department of Geography and Geoinformation Science, George Mason University, Fairfax, VA, USA; [b]Center for Geoinformatics and Geospatial Intelligence, George Mason University, Fairfax, VA, USA; [c]Center for Earth Observing and Space Research, College of Science, George Mason University, Fairfax, VA, USA; [d]Criminal Investigations and Network Analysis Center, George Mason University, Fairfax, VA, USA; [e]Department of Computational and Data Sciences, George Mason University, Fairfax, VA, USA

ABSTRACT

The worldwide slum population currently stands at over one billion, with substantial growth expected in the coming decades. Traditionally, slums have been mapped using information derived mainly from either physical indicators using remote sensing data, or socio-economic indicators using census data. Each data source on its own provides only a partial view of slums, an issue further compounded by data poverty in less-developed countries. To overcome such issues, this paper explores the fusion of traditional with emerging open data sources and data mining tools to identify additional indicators that can be used to detect and map the presence of slums, map their footprint, and map their evolution. Towards this goal, we develop an indicator database for slums using open sources of physical and socio-economic data that can be used to characterize slum settlements. Using this database, we then leverage data mining techniques to identify the most suitable combination of these indicators for mapping slums. Using three cities in Kenya as test cases, results show that the fusion of these data can improve the mapping accuracy of slums. These results suggest that the proposed approach can provide a viable solution to the emerging challenge of monitoring the growth of slums.

ARTICLE HISTORY
Received 10 July 2018
Accepted 26 November 2018

KEYWORDS
Slums; remote sensing; socio-economic; urban sustainability; data mining; Kenya

1. Introduction

Cities worldwide are presently confronted with the challenge of having to meet the needs of a constantly growing population (Bhatta 2009; Pramanik and Stathakis 2016). While developed countries tend to have the necessary infrastructure to accommodate this growth, many less-developed countries are facing insurmountable social, economic and developmental challenges as a result of this trend (Cohen 2006; Hassan and Mahabir 2018). One of the consequences of this challenge in less-developed countries is the emergence of slums, which represents a large portion of the urban population. It is currently estimated that nearly one billion people worldwide live in slum settlements. This number is projected to increase to 2 billion by 2030 and 3 billion by 2050 if current trends persist (United Nations 2015).

Meeting the growing demands of increasing urban population while at the same time ensuring their sustainability (e.g. Sustainable Development Goals – UNDP 2015) requires the continual collection of information on slum growth and their conditions. Such information includes the number

CONTACT Ron Mahabir rmahabir@gmu.edu

of people living in slums, their growth/decline, and any changes to their physical characteristics (e.g. durability of housing) and socio-economic characteristics (e.g. adequate access to water and sanitation) over time. Similar to work on cities at large (e.g. Craglia et al. 2004; Hollander et al. 2017), such information can be conveyed in the form of various indicators, which can be used to monitor and assess the different characteristics of slums. This information can then be used to determine whether the policies, programs and resources put in place to improve the conditions in slums, and the quality of life of slum dwellers are effective.

As it relates to this work, indicators represent specific, observable, and measurable characteristics that can be used to track changes in a particular phenomenon over time (readers interested in a more in-depth review of indicators and their role in the urban setting are referred to Godin (2003) and Kitchin, Lauriault, and McArdle (2015) for further discussion). While various indicators are available for monitoring different aspects of cities (e.g. World Council on City Data (2017) – WCCD), these indicators are often unreliable or unavailable when it comes to slums for various reasons. For example, population census and household surveys may be difficult or even impossible to collect or access in some less-developed countries (Mahabir et al. 2016). To overcome such issues, remote sensing often offers an alternative source of information. However, due to the nature of remote sensing data, it is possible to directly capture only the physical characteristics of the slum, missing the equally important social component. The recent emergence of open data presents new opportunities to mitigate these possible limitations (GFDRR 2014). Not only does it offer an alternative and free or low-cost source of information on slums, but such data can also be complementary to traditional sources of slum data (Chakraborty et al. 2015), and further, potentially allow for a suite of slum indicators to emerge. However, the use of open data for mapping and monitoring slums has yet to be fully exploited (Chakraborty et al. 2015).

While open data has the potential to serve as a rich source of information for understanding slums, harnessing it for this purpose presents some newfound challenges. Specifically, the sheer volume, variety, and velocity of such open data (Crooks et al. 2015; Schaffers et al. 2011) requires the capacity to efficiently sift through large and diverse datasets to identify the most suitable data elements for deriving indicators, for monitoring and assessing slums. Data mining approaches can often offer solutions to help sieve through large amounts of data in order to identify application-specific indicators. Although such approaches have been used to study slums, there has been limited work in this area, with most studies using remote sensing imagery alone to study slums (see, Kuffer, Pfeffer, and Sliuzas (2016) and Mahabir et al. (2018a) for a review of related work in this area). Very few studies have combined multiple socio-economic and physical indicators of slums to map them. One notable exception is that of Hacker et al. (2013), which combined both land cover data derived from Landsat imagery and socio-economic data derived from census data to map slums.

Motivated by these challenges, this paper explores how open data can be used as content-rich alternative sources of information to complement remote sensing imagery in supporting slum detection, mapping and monitoring. Accordingly, the research presented here pursues a twofold objective. First, it aims to identify which open data sources can be combined with remote sensing data to identify and map slums in data-poor areas where the problem of slums is most acute. Secondly, given these newly created data sources, it aims to identify to most suitable indicators to map and monitor slums. In order to accomplish this, data mining techniques are used to develop context-sensitive definitions for slums based on location, as well as for testing the generalizability of indicators and derived slum models. By addressing these objectives, we demonstrate not only that each slum has unique physical and socio-economic qualities, but also that we can use data mining approaches to extract these indicators for mapping and monitoring them.

The remainder of this paper is structured as follows. In Section 2, we review issues with current slum indicators and the need for pursuing opportunities using open sources of data for acquiring information on slums. This section also briefly reviews the topic of data mining and its potential for combining slum indicators towards the identification and mapping of slums. Section 3 presents

the study sites and slum data used in this research. Section 4 discusses the approach used to identify relevant slum indicators and their application in the mapping of slums. Following, an analysis and interpretation of our findings are reported in Section 5. In Section 6, a discussion and conclusion summarizing this paper, along with recommendations for future work are presented.

2. Background

For as long as cities have been around, slums have existed with their presence long been documented (e.g. Booth 1903). The study of slums has mainly followed one of three lines of enquiry: (1) socio-economic and policy issues, (2) physical aspects, and (3) modeling their dynamics and growth (Mahabir et al. 2016). Most typical indicators used to study slums, however, are derived from data collected on physical and socio-economic aspects of slums. For example, the United Nations Human Settlement programme (UN Habitat 2006) identifies slum households as lacking one or more of the following characteristics:

(1) Durable housing of a permanent nature that protects against extreme climate conditions,
(2) Adequate living space (no more than three people sharing the same room),
(3) Easy access to safe water in sufficient amounts at an affordable price,
(4) Adequate access to sanitation in the form of a private or public toilet shared by a reasonable number of people, and
(5) Security of tenure that prevents forced evictions.

Using this definition, one can develop physical indicators (the first characteristic) and socio-economic indicators (the remaining four characteristics) that define a slum. However, this global definition may fail to capture the local perceptions of slums. For example, UN Habitat (2003) discussed how many countries have their own unique combination of indicators pertaining to what is a slum, which may include, crime and violence, health/hygiene, and construction legality, illustrating the inherent fuzziness when it comes to defining a slum. Patel, Koizumi, and Crooks (2014), using slums in India as a case study, further show that even within the same country slum definitions can vary among different levels of governance.

This problem is further compounded by challenges in collecting relevant data on slums. Information on security of tenure, for instance, is limited in many countries and is therefore not included in any global estimates of slum populations (UNSD 2012). In this context, the limited coverage and availability of data over slum areas, which effectively makes them data deserts (Zetter and De Souza 2000), introduces an additional level of difficulty. For example, while survey-type data such as population census and household surveys have been widely used for deriving socio-economic indicators (e.g. Baud, Sridharan, and Pfeffer 2008; Nolan 2015; Weeks et al. 2007), such data are usually released for public use at coarse enumerated levels (e.g. city or state). In addition, as census data collection occurs at time intervals of several years (or decades, as in the case of Myanmar (UNFPA 2018) or Lebanon (Kattan et al. 2016)), they often fail to capture the dynamic nature of slums. This is particularly important in fast-growing slums, such as in Dhaka, Bangladesh where population increase can reach upwards of 1000 people per day (Kotkin 2014). This is in comparison to daily estimates of 1200 people per day if all population groups (including slums) are considered (Taubenböck and Wurm 2015). While approaches exist for overcoming such temporal mismatches in the collection of census data (e.g. regression and interpolation techniques), the application of these techniques over very large areas may lead to loss of more subtle – yet important – changes for relatively smaller geographic features such as slums. Consequently, census data alone is insufficient for capturing the spatial and temporal variation necessary to map and monitor slums.

Compared to socio-economic indicators, physical indicators of slums (e.g. high dwelling density, small size and location on hazardous land) have mainly been derived from remote sensing data (e.g. Owen and Wong 2013; Wurm and Taubenböck 2018). Despite its well-known advantages, remote

sensing of slums is not without challenges. Specifically, a key challenge is the difficulty in inferring socio-cultural and socio-economic information from remote sensing data alone, as it assumes a direct link between physical properties and such information (Angeles et al. 2009). Such a link is often obscure, as, for instance, in the Paraisópolis slum in San Paulo, Brazil, which has land divisions like that of formal settlements (Saraiva and Marques 2004). If observed only through remote sensing imagery, these slums may be misidentified as formal settlements.

The emergence of open data provides a newfound opportunity to supplement these more traditional approaches to data collection and access (Crooks et al. 2016; Logan et al. 2017). Today, governments (e.g. Data.gov in the United States), non-government organizations (e.g. Data.worldbank.org by the World Bank), and industry (e.g. google.com/mapmaker by Google) are sharing their data openly, frequently by leveraging web technologies. While many such initiatives have traditionally been centered in developed countries, they are becoming prevalent in less-developed countries as well. The Kenya Open Data initiative (KOD 2018) is one such example. Kenya Open Data provides datasets ranging from infrastructure (e.g. roads), to services (e.g. water and sanitation), and socio-economic conditions (e.g. income levels). Similar to such national-scale efforts, several regional open data initiatives have emerged in the less-developed countries. Such initiatives include Afrobarometer (2018) and openAFRICA (2018) in Africa, the Latin American Open Data Initiative (2018) in Latin America, and Open MENA (2018), which covers Middle Eastern and North African regions. These platforms provide users access to free and diverse data that can be leveraged to narrow down existing data gaps common in many less-developed countries.

A prominent example of such a data source is OpenStreetMap (OSM), a freely accessible and editable map of the world (Haklay 2010; Jackson et al. 2013; Mooney and Cocoran 2014; Mullen et al. 2015). Open data platforms such as OSM can provide highly detailed information on slums such as the locations of medical clinics, food stalls, and schools, which may not otherwise be available. Mahabir et al. (2017), for example, showed that some slums in Nairobi, Kenya were better served by open sources of road data acquired from OSM compared to data acquired from the available authoritative source. Furthermore, this data has also been studied in the context of the wider news awareness ecosystem, in order to better understand the complex links between news awareness and digital activism (Mahabir et al. 2018b). Moreover, it has also been shown that open data, in conjunction with social media sources can be used to characterize and distinguish between different human activities with a city (e.g. Jenkins et al. 2016). Thus, in addition to the mapping and monitoring of slums, open data may provide opportunities for capturing new insights on slums and the people that live in them.

While open data and its integration with more traditional authoritative data sources offer a promising avenue for the derivation of indicators, it also presents some emerging challenges. Chief of these is the issue of integrating data at different scales and granularities (Openshaw 1983; Robinson 1950). For example, fine spatial resolution remote sensing data may need to be combined with data available at a much coarser level of detail (e.g. census surveys). In such cases, a data layer may need to be converted from one spatial scale to another (e.g. aggregating to a coarser spatial resolution) so that it can be compatible for combination with other data. This data conversion process often results in a converted dataset that is of lower quality than the data from which it was derived (Gotaway and Young 2002). Notwithstanding, studies have shown that combining indicators captured at different scales can be complementary with respect to improving modeling/classification results (e.g. Bhaduri et al. 2007; Grant, Gennings, and Wheeler 2015; Rupert 2003; Stevens et al. 2015). A recent study by Steele et al. (2017), for example, utilized mobile and remote sensing data captured at different scales for Bangladesh. This data was modeled against a derived wealth index (WI) constructed using data from the Demographic and Health Survey for that country, showing a good relationship between these indicators. That study further showed that the mean WI values extracted for slums can be used to distinguish slums from formal settlements. However, the focus of that study was not the mapping of slums. Few studies have examined this specific issue in the context of using multiple indicators captured at different scales, along with understanding the characteristics of such indicators that make them more or less suitable for mapping slums, which this study is addressing. This is

important since different factors (as represented in data and derived indicators) help shape the formation and growth of slums (Patel, Crooks, and Koizumi 2012).

Another significant challenge related to the greater availability and variety of slum-related data is the selection of a set of indicators that is most suitable in the context of each specific slum. This has led to variations in the set indicators that are considered suitable for mapping slums by different experts (e.g. Kohli et al. 2012). One data-driven approach to overcome this challenge is to derive indicators from large amounts of curated data of different types and at different spatial scales using data mining. Formally defined, data mining is 'the non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data' (Fayyad, Piatetsky-Shapiro, and Smyth 1996). Various efforts have been made to use data mining to explore various aspects of slums. For example, Ribeiro (2015) used a C4.5 decision tree (Quinlan 1996) to map different land cover classes in Embu, Brazil. The land cover classes were then used to determine different land uses, including slums. That study reported producer's and user's classification accuracies for land cover classes upwards of 73%. Similar work using decision trees include that of Owen and Wong (2013) in Guatemala, Du, Zhang, and Zhang (2015) in China, and Kuffer et al. (2016) in India and Rwanda, all of which report overall classification accuracies of 80% or more using such a method.

While the research noted above focused primarily on optical imagery to map slums, few studies have also explored the use of active sensors. For example, Wurm et al. (2017) applied two different methods, a random forest tree classier and a linear discriminant analysis classifier, to map slums in radar imagery over Mumbai, India. Various image-based textures and morphological profile indicators were considered. These were captured at different scales, with the largest window size of 81×81 pixels leading to the highest overall classification accuracy of 89%.

In addition to the use of decision trees, other data mining techniques have also been used to study slums. Dubovyk, Sliuzas, and Flacke (2011) applied logistic regression models to study slums in Istanbul, Turkey. By using various proximity, site-specific and neighborhood factors responsible for the development of existing slums, they were able to predict new slums areas. Khelifa and Mimoun (2012) used a genetic algorithm to define rules for slum and formal settlements following the segmentation of high-resolution imagery over Oran city in Algeria. Other studies, such as Ali, Hegazy, and Eldien (2010), further applied Neural Networks (NNs) to slums, while Dahmani, Fora, and Sbihi (2014) and Vatsavai (2013), used Support Vector Machines (SVMs) and multiple instance learning (MIL) approaches to study slums. A common thread among these studies is the use high and very high-resolution (H/VH-R) remote sensing data. However, such data, while useful in capturing a finer grain view of slums, may not always be accessible for some less-developed countries due to the relatively high cost of such data. Likewise, very few studies have applied data mining to more than one site and/or for large geographic areas (e.g. an entire metropolitan area).

Another issue with H/VH-R imagery is that for longitudinal studies, this data only captures more recent physical changes that have occurred within slums. For some slums, such as Kibera, examined in our study; this slum was set up in the early 1900s (Parsons 1997). By the early 1990s, Kibera's population had already reached 250,000 people, and its internal and external morphology has changed substantially (Government of Kenya 2001). In such cases, Landsat imagery, with an archive dating back to the early 1970s, may be the only feasible source of earth observation data for tracking the growth and evolution of these slums. As the availability of longitudinal data via H/VH-R image sources continue to grow, this data will potentially become an alternative to longitudinal Landsat data for monitoring and tracking the growth of slums. Such data is also expected to be complemented by a slew of other more recent emerging sources of imagery data, including, newer Sentinel missions, Lidar and unmanned aerial vehicles.

Given these considerations, this paper aims to examine how various sources of remote sensing, together with the recent increase in availability and richness of open socio-economic data, can be leveraged towards deriving indicators for identifying and mapping slums over large geographic areas. Specifically, we aim to show how new slum-related indicators can be identified and derived

through a data-driven approach that combines the richness of open data with the effectiveness of data mining. In line with this, and due to the unique physical and socio-economic characteristics typical of slums, indicator thresholds are expected to vary across slums. As such, a major component of this research also involved the identification indicator thresholds for slums at each study location in a data-driven manner. In so doing, while the indicators for mapping slums may vary from one study area to the next, a benefit of the methodology presented here is that it can be used to better understand how these two properties of slum indicators (i.e. the specific indicators and their individual thresholds) change over time.

3. Study areas and data

For this study, three sites which represent different geographies in Kenya were selected, specifically the cities of Nairobi, Mombasa and Kisumu. These sites were selected for two primary reasons. First, they represent the largest cities in Kenya with the highest population of slum dwellers. Secondly, ground truth data on slums, which is required for assessing the performance of indicators, is available. Figure 1 shows the locations of all three study areas in relation to Kenya. Data on the spatial extent of slums were collected from MajiData (2016), which was set up by the Water Services Trust Fund in Kenya. This data resulted from a corporation established between Kenya's Ministry of Water and Irrigation, the Kenyan Government, various international donors, and community stakeholders. Based on data availability, we considered a study period of 5 years between 2010 and 2014.

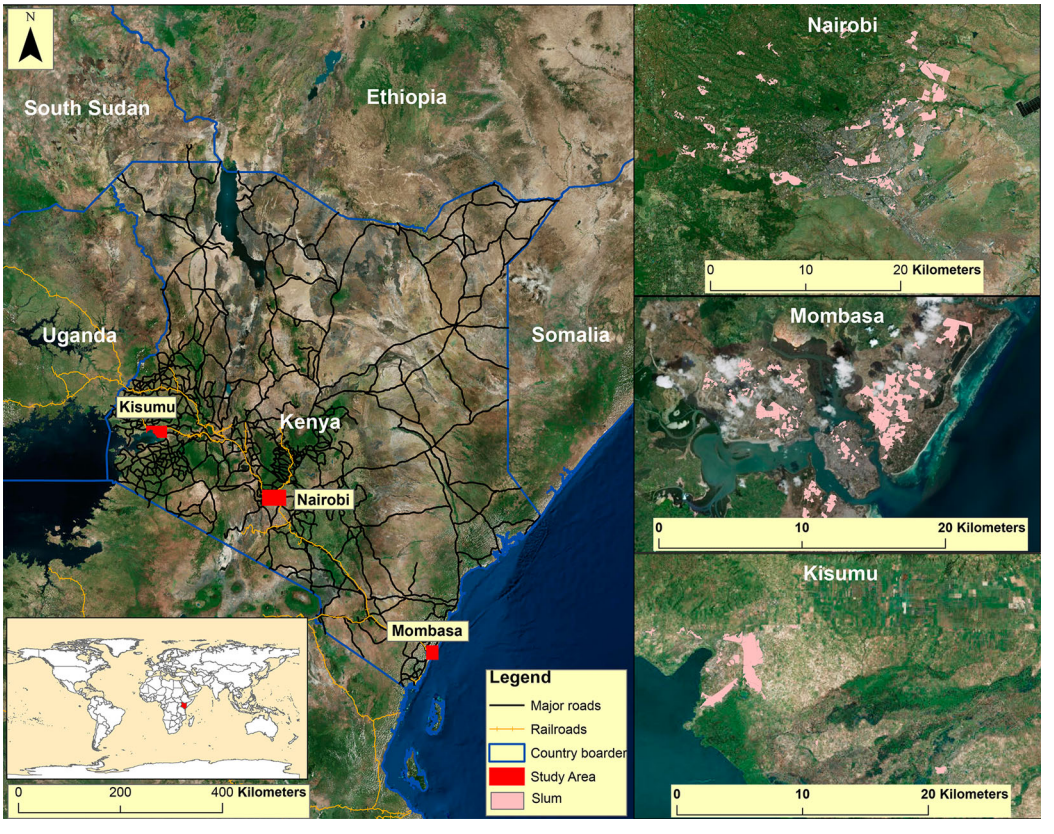

Figure 1. Study areas in Kenya.

4. Methodology

In order to derive slum indicators using a data-driven approach, the methodology used in this research is based on five consecutive steps. First, we define a set of requirements for candidate indicators that enable their comparison and assessment between cities (Section 4.1.). Based on this definition, a pool of candidate indicators is extracted from the various data sources (Section 4.2.). Candidate indicators are further processed (Section 4.3.) and used to create a set of models for detecting and mapping slums (Section 4.4.). Finally, each model is validated against ground truth data (Section 4.5.). These steps are described in the sections that follow with the overall methodology visualized in Figure 2.

4.1. Defining candidate indicator requirements

In the context of this work, any measure that is considered to be a possible indicator ('candidate indicator') must fulfill the following four basic requirements: (1), the candidate indicators should be based entirely on open data to overcome the data poverty challenges previously discussed; (2), the candidate indicators should be quantitative and in a form that lends itself to subsequent spatial analysis; (3), the candidate indicators should be captured across the entire area of interest; and (4), the candidate indicators should be available for all three study areas such that a comparison of indicators across the study areas can be performed. Moreover, it is required that the candidate indicators would be available for the study period and that they could be systematically collected on a recurring basis of every three to five years (or more frequent if possible). This last requirement specification was used to ensure that derived indicators could be collected on a frequent basis, albeit every several years, which would, therefore, allow the monitoring of slums over time.

4.2. Identification and extraction of candidate indicators

Based on the UN Habitat (2006) definition of slums that was outlined in Section 2, and given the requirements for candidate indicators outlined above, candidate indicators were divided into two broad categories: physical and socio-economic. Physical indicators provide information on the morphological characteristics of an area, while socio-economic indicators provide underlying information, directly or indirectly, on the population. Additionally, candidate indicators were labeled as primary or auxiliary, depending on the elements they capture in the UN Habitat (2006) definition or in the wider slum literature reviewed here.

A summary of the candidate indicators, including their data source, the corresponding hypothesis (for including them) and required processing steps, is provided in Table 1. In total, six thematic classes of candidate indicators were identified: (1) durable housing, (2) adequate living space,

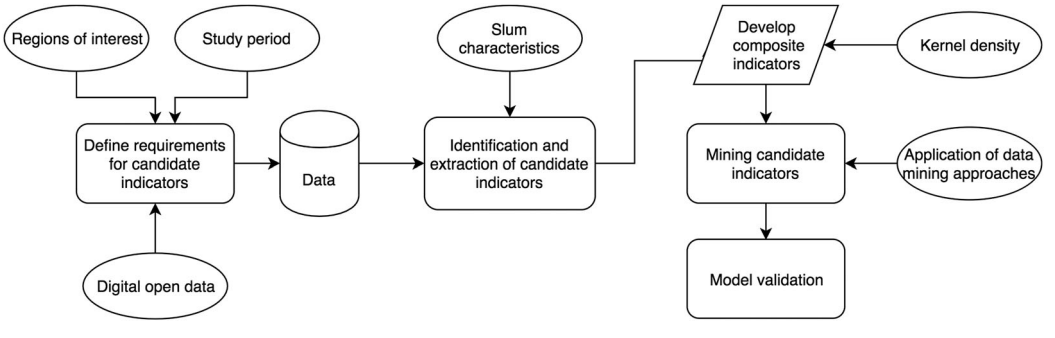

Figure 2. Methodology workflow.

Table 1. Candidate indicators for slums.

Candidate indicator group (number of indicators in group)	Data source	Description (Year/period)	Hypothesis	Processing steps
Class 1: Durable housing – UN Habitat requirement				
Hazardous locations (1)	OpenStreetMap (2017) Google Map Maker (2017)	Polyline vector (2014) Polyline vector (2014)	Slums are more likely to be in closer proximity to major roadways.	1. Primary and secondary roads from each data source were extracted and merged into a single dataset.
Image texture (48)	Landsat 8 panchromatic band – Annual median composite (Google Earth Engine 2017)	15 m spatial resolution (2014)	Certain texture measures at different scales (window sizes) will provide good separability between slums and other non-slum areas.	1. Various texture measures were extracted: contrast, correlation, dissimilarity, entropy, homogeneity, mean, second moment and variance. A unit shift value of 1 pixel in both *x* and *y* directions was used for these calculations. 2. Each texture measure was extracted at different window sizes: 3×3, 5×5, 7×7, 9×9, 11×11 and 13×13. This resulted in a total of 48 image texture measures.
Raw image bands (9)	Landsat 8 – Image bands 1–7 and 10–11 Annual median composite (Google Earth Engine 2017)	1. Image bands 1 –7, 30 m spatial resolution (2014) 2. Image bands 10–11, 100 m spatial resolution (2014)	Certain raw image bands are more sensitive to slums than others.	1. Nine raw image band indicators were computed.
Class 2: Adequate living space – UN Habitat requirement				
Population density (1)	LandScan (ORNL 2017)	Raster image with 1 km by 1 km spatial resolution (2014)	Slums have higher population densities compared to other types of communities.	1. Population density was extracted from the dataset for each $1\text{ km}\times1\text{ km}$ grid cell.
Birth rates(1)	WorldPop birth rates (WorldPop 2017)	Raster image with 100 m by 100 m spatial resolution (2015)	Slums have higher birth rates compared to other types of communities.	1. Birth rates were extracted from the dataset for each $100\text{ m}\times100\text{ m}$ grid cell.
Class 3: Access to safe water – UN Habitat requirement				
Water kiosks (1)	OpenStreetMap (2017) Google Map Maker (2017)	Point vector (updated until 2014) Point vector (updated until 2014)	Slums have high a large number of water kiosks.	1. The XY location of water kiosks was extracted from both data sources and merged into a single dataset.
Class 4: Access to adequate Sanitation – UN Habitat requirement				
Pit latrines (1)	OpenStreetMap (2017) Google Map Maker (2017)	Point vector (updated until 2014) Point vector (updated until 2014)	Slums have a large number of pit latrines.	1. The XY location of pit latrines was extracted from both data sources and merged into a single dataset.
Class 5: Security of tenure – UN Habitat requirement				
Online real estate activity (1)	Online real estate agencies in Kenya and Africa (e.g. Property24	XY location and attribute characteristics of online real estate activity –	Slums, due to their typical illegal occupancy of the land	1. The XY locations of real estate transaction locations were

(*Continued*)

Table 1. Continued.

Candidate indicator group (number of indicators in group)	Data source	Description (Year/period)	Hypothesis	Processing steps
	– www.property24.co.ke)	rental, sale or purchase of house, commercial property, building or land – (2010–2014)	in which they are built, have little to no online real estate activity.	collected from various online websites and merged into a single dataset.
Class 6: Auxiliary Candidate Indicators				
Quality of life – NDVI (1)	Landsat 8 – Image bands 4 and 5 (Google Earth Engine 2017)	30 m spatial resolution (2014)	Slums have low amounts of vegetation compared to other settlement types.	1. This dataset was created by generating an NDVI layer for each set of Landsat imagery collected for the year 2014. The median value for each pixel across this stack of NDVI images was then extracted to create a greenest pixel NDVI dataset. 2. Greenest pixels values were rescaled to values between 0 and 1 to have only positive instances of this indicator.
Roads – poor quality (1)	OpenStreetMap (2017) Google Map Maker (2017)	Polyline vector (2014) Polyline vector (2014)	Slums have a high quantity of poor road types.	1. Poor quality roads (i.e. unpaved, unsurfaced, track, path, footpath, gravel) were extracted from each data source and merged into a single dataset.
Roads – dead ends (1)	OpenStreetMap (2017) Google Map Maker (2017)	Polyline vector (2014) Polyline vector (2014)	Slums have a high quantity of dead-end streets.	1. Road data for both datasets were merged into a single layer. 2. Nodes in this street layer with only one connected edge were considered dead-end nodes. These nodes were extracted.
Roads – average curvature (1)	OpenStreetMap (2017) Google Map Maker (2017)	Polyline vector (2014) Polyline vector (2014)	Roads in slums have high average curvature values (due to poor urban planning practices) indicating more irregular roads.	1. An algorithm was written to examine the vector angle between every 3 nodes in the road layer. For example, if a segment of road contained 5 nodes (n1, n2, n3, n4 and n5), then angles were calculated as follows: n1, n2, n3; n2, n3, n4; n3, n4, n5. 2. The average of these angles was then computed for each road segment, where a road segment was represented as the segment between two intersection nodes, one

(*Continued*)

Table 1. Continued.

Candidate indicator group (number of indicators in group)	Data source	Description (Year/period)	Hypothesis	Processing steps
				intersection and one dead-end node, or two end nodes.
Roads – total length per unit area (1)	OpenStreetMap (2017) Google Map Maker (2017)	Polyline vector (2014) Polyline vector (2014)	Slums have lower road densities per unit area when compared to other settlement types.	1. The total length of roads from the two data sources was merged and computed for each 1 km by 1 km grid cell.
Road – roads per unit population (1)	Road – OpenStreetMap (2017) Road – Google Map Maker (2017) Population – LandScan (ORNL 2017)	Polyline vector (2014) Polyline vector (2014) 1 km by 1 km spatial resolution (2014)	Slums have lower amounts of roads per unit population.	1. The total length of roads per 1 km × 1 km grid cell were divided by the population density contained within that grid cell.
Services – private primary schools (1)	Kenya Open Data (KOD 2018)	Point vector (updated until 2014)	A large number of private primary schools are located in and immediately around slums.	1. XY locations of private primary schools were extracted from the data source.
Services – private health care facilities (1)	Kenya Open Data (KOD 2018)	Point vector (updated until 2014)	A large number of private health care facilities are located in and immediately around slums.	1. XY locations of private health care facilities were extracted from the data source
Services – places of worship (1)	OpenStreetMap (2017) Google Map Maker (2017) Kenya Open Data (KOD 2018)	Point vector (updated until 2014) Point vector (updated until 2014) Point vector (updated until 2014)	A large number of places of worship are located in and immediately around slums.	1. Places of worship (e.g. churches, masjids and temples) were extracted from each data source and combined.
Ambient Geospatial Information – Flickr (1)	Flickr API (Flickr 2017)	Location of Flickr images containing the keywords 'slum' or 'informal settlement' in their tags (2010–2014).	A large number of geolocated Flickr points containing the keywords 'slum' or 'informal settlement' in their tags are clustered in and around slums.	1. A script was written to query the Flickr API using search terms 'slum' or 'informal settlement' 2. The XY locations of features resulting from this search query were then merged into a single dataset.
Local news media – text mining (1)	Local online news media sites in Kenya (e.g. Daily Nation – https://www.nation.co.ke)	Online news articles discussing slums in Kenya (2010–2014).	A large number of news media articles on slum settlements are based on reported activity within and around slums.	1. Articles from Google News were curated using the search terms 'slum' or 'informal settlement'. 2. A script was written to then extract the sentence in each script containing these search terms. The overall assumptions were that the name of the slum would be likely to occur within this sentence of words. 3. Sentences were then individually tokenized, their nouns and proper

(*Continued*)

Table 1. Continued.

Candidate indicator group (number of indicators in group)	Data source	Description (Year/period)	Hypothesis	Processing steps
				nouns extracted, and these further geocoded using both OSM and Google Geocoding services. 4. Geocoding was restricted to the specific study areas to reduce the instance of false positives. 5. The locations of features resulting from successful geocoded names were then combined into a single dataset.

(3) access to safe water, (4) access to adequate sanitation, (5) security of tenure, and (6) auxiliary candidate indicators. Within each class, candidate indicators were grouped based on the data and the processing steps that were used to derive them. The first five classes of indicators in Table 1 are based on the UN Habitat (2006) slum requirements and utilize data collected from various vector and raster open data sources as described in column 2 of the table. The remaining columns – columns 3, 4 and 5 in Table 1 – describe the period covered by each data set, the hypothesis for including the candidate indicator group in our analysis, and the processing steps used to derive the candidate indicators, respectively.

Overall, the first five classes include 63 candidate indicators. It is important to note that several of the candidate indicators were derived from multispectral imagery, resulting in multiple candidate indicators. For example, there were 48 textural indicators derived from the panchromatic band of the Landsat imagery used. Our use of this image band was guided by previous work (Lu and Weng 2005), which showed that for this specific image data source, image texture derived from the panchromatic band preserved much more of the contextual details of the underlying land use and land cover compare to other image bands.

The sixth class, auxiliary candidate indicators, represent additional properties of slums not directly outlined in the UN Habitat (2006) requirements. Overall, 11 candidate auxiliary indicators were identified in this class. This class was included in our study in order to build a more comprehensive pool of candidate indicators that complement those derived for UN Habitat (2006), which may provide a higher potential for generalization of the indicators. It is worth noting that the data specified in Table 1 may vary in its spatial scale, which may lead to some inconsistencies at the slum dwelling level. However, as we aim to explore the use of open data at settlement-level rather than the dwelling level, such possible inconsistencies are not likely to alter the outcomes of this research.

4.3. Data processing

In order to facilitate the process of data-driven composite indicator discovery, the candidate indicators specified in Table 1 were left unchanged with the exception of three cases: water and sanitation, urban services, and geosocial media. The rational for doing this was that while most candidate indicators are based on a systematic sampling process (e.g. remote sensing coverage), this was not the case for these three cases. In the Kibera slum in Nairobi, for instance, the presence

of water kiosks may be disproportionately skewed due to the greater presence of non-government organizations (e.g. United Nations) in and around this slum. In that respect, adequate sanitation is similar in the sense that its coverage may be skewed to concentrated efforts by humanitarian projects, or other similar endeavors. In the case of safe water and adequate sanitation, we, therefore, combined these candidate indicators as they are often discussed in unison in the literature (e.g. Isunju et al. 2011). The cases of urban social services (private primary schools, private health clinics and places of worship) and geosocial media (Flickr and local news media) are similar to water and sanitation in that there is a natural relationship between the various candidate indicators that were combined. As a result, a new revised list of 69 candidate indicators was identified for detecting and mapping slums.

Following the revision of candidate indicators, the data type of some indicators had to be converted from vector to raster using the kernel density method with a bandwidth of 1 km, and a cell size of 15 m. Additionally, in order to enable a comparison between candidate indicators and the ground truth data, the vector data containing the spatial extent of slums (ground truth data) for each study area were also converted into 15 m by 15 m binary grid of slum and non-slum cells.

4.4. Mining candidate indicators

After the candidate indicators were derived and computed, several data mining approaches are used to test the suitability of these indicators for mapping slums. Specifically, these were logistic regression (Cox 1958), discriminant analysis (Fisher 1936), and the See5 decision tree (Quinlan 2015). These approaches were selected specifically for their ability to provide information that can be used for assessing the suitability of each indicator for identifying and mapping slums. Furthermore, these approaches have been widely accepted in the data mining community (Olson and Delen 2008; Wu et al. 2008), and have been implemented in various open source software and programming packages (e.g. Weka (Frank, Hall and Witten 2016), the R Project (R Core Team 2018) and the Scikit library (Pedregosa et al. 2011) in Python). These three approaches have also been used independently for the detection and mapping of slums (e.g. Dubovyk, Sliuzas, and Flacke 2011; Graesser et al. 2012; Kuffer et al. 2017; Owen and Wong 2013). However, to the authors' knowledge, these approaches have not been systematically compared and contrasted for the study of slums.

4.5. Model validation

To assess the classification performance of each data mining approach, precision and recall accuracy measures were calculated (e.g. Pratomo et al. 2017). This was done using a five-fold cross-validation approach to account for the very skewed distribution of non-slum cases compared to positive cases (this issue is discussed in the next section) and using the slum data sourced from MajiData (2016). Moreover, in order to test the generalizability of the models, models developed for one study area were then applied to the other two study areas and their classification accuracies assessed.

5. Results and analysis

The methodology outlined above was applied to the three case study sites. Table 2 provides a summary of the distribution of classes across all three sites. As can be seen from this table, the rates of

Table 2. Distribution of slums cases.

Study area	Slum cases (% of total)	Non-slum cases (% of total)	Total number of cases
Nairobi	3.40	96.60	1,205,728
Mombasa	8.48	91.52	1,048,407
Kisumu	1.50	98.50	717,545

Table 3. Overall classification accuracies.

	Nairobi		Mombasa		Kisumu	
Model	Precision	Recall	Precision	Recall	Precision	Recall
LR	0.01	0.10	0.29	0.57	0.45	0.72
DA	0.66	0.22	0.79	0.35	0.71	0.36
See5	0.90	0.98	0.93	0.95	0.93	0.94

slum cells are low (between 1% and 9%) in comparison to the rates of non-slum cells. The classification results of each model (logistic regression, discriminant analysis and See5 decision tree) are summarized in Table 3. This table reports the precision and recall measures, which provide an assessment of the overall classification accuracy. A comparison of the accuracies across all three sites shows that the See5 decision tree consistently outperforms the other approaches. The spatial footprint of the slum cell classification results for each model (for the three case studies) using the ground truth data are shown in Figure 3.

In order to gain a better understanding of these results, and in particular the contribution of specific composite indicators to the classification accuracy, as well as the overall generalizability of the proposed approach, our discussion of the results is organized as follows. Sections 5.1.–5.3. (*Logistic regression, Discriminant analysis and See5*) provide a more thorough discussion of the classification results, with an emphasis on the behavior of each model in the three case studies.

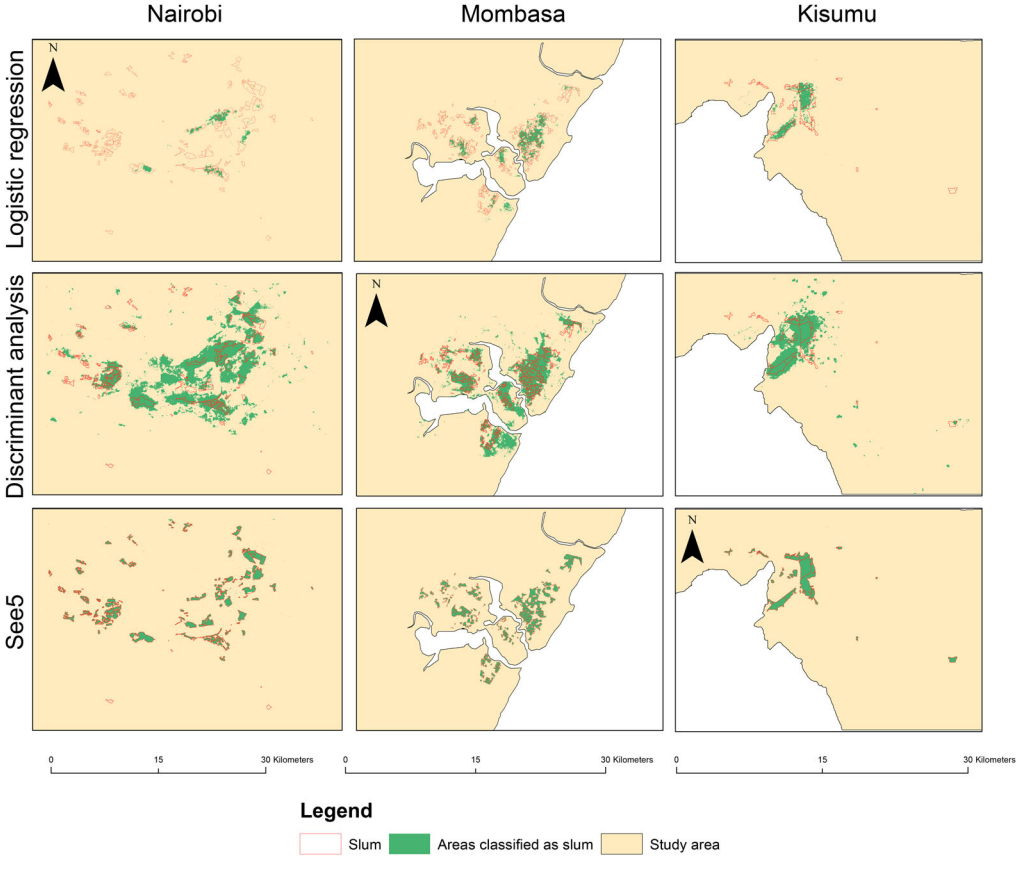

Figure 3. Distribution of positive classified cases for slums for (a) logistic regression, (b) discriminant analysis and (c) the See5 decision tree.

Following this, we turn our attention to additional analyses that were performed in order to understand the contribution of specific indicators to the overall classification accuracy (Section 5.4) and to test the generalizability of slum data models (Section 5.5).

5.1. Logistic regression

The results in Table 4 show that the logistic regression (LR) model performed poorly with low precision values ranging between 0.01 and 0.45 for all study sites. These values reflect the low number of correctly classified slum cells and the large number of false positives (type I errors). In Nairobi, as shown in Figure 3(a), most cells classified as slums were located in and around the immediate vicinity of the larger slums, which were also close to the downtown Nairobi district. Areas immediately surrounding these larger slums share similar physical slum-like characteristics, for instance, a high density of small rooftops, which helps to explain the misclassification of these areas as slums. Most smaller slums were not detected using this classification approach. The Nairobi site also had very low recall values indicating a high percentage of false negatives (type II errors) in the LR model results for this site. The large number of type I and II errors reported for Nairobi highlight the underlying difficulty of LR model to find suitable thresholds for separating between the slums and non-slum cells using the derived indicators. This is in part due to the complexity of this site, with many slums having characteristics similar to that of formal settlements.

Notably, in Mombasa, the spatial coverage of correctly classified slum cells was substantially more dispersed compared to Nairobi. This could be in part due to the more uniform distribution of slum sizes based on the ground truth data – which would, in turn, lead to less skewed results. Slums in the northwestern areas had little to no coverage with many type I errors occurring around the boarders of slums. The precision values for Mombasa were also relatively high (as shown in Table 2), indicating that the LR model for this site performed significantly better than the LR model for Nairobi.

In Kisumu, classification accuracy increased further with respect to both precision and recall metrics compared to the other two sites. However, coverage was mainly restricted to the few large

Table 4. Indicators for first four levels for See5 tree for each study site.

	Nairobi					Mombasa					Kisumu					*All sites*
	Tree level					Tree level					Tree level					
Candidate indicators	1	2	3	4	Total	1	2	3	4	Total	1	2	3	4	Total	*Total*
Socio-economic																
Average road curvature	0	0	0	0	**0**	0	0	1	2	**3**	1	0	1	0	**2**	*5*
Births	1	0	0	0	**1**	0	0	0	0	**0**	0	1	0	0	**1**	*2*
Dead ends	0	0	0	1	**1**	0	0	0	0	**0**	0	0	0	4	**4**	*5*
Population	0	0	2	0	**2**	0	0	0	0	**0**	0	0	0	0	**0**	*2*
Real estate activity	0	0	0	0	**0**	0	0	1	1	**2**	0	0	2	0	**2**	*4*
GeoSocial media	0	0	1	0	**1**	0	0	0	1	**1**	0	0	1	0	**1**	*3*
Services	0	1	0	2	**3**	1	0	0	1	**2**	0	1	0	0	**1**	*6*
Total per level per site	*1*	*1*	*3*	*3*	***8***	*1*	*0*	*2*	*5*	***8***	*1*	*2*	*4*	*4*	***11***	*27*
Satellite-derived indicators																
Greennest	0	0	0	1	**1**	0	0	1	0	**1**	0	0	0	0	**0**	*2*
Entropy 13 × 13	0	0	0	0	**0**	0	0	0	1	**1**	0	0	0	0	**0**	*1*
Second moment 9 × 9	0	0	0	1	**1**	0	0	0	0	**0**	0	0	0	0	**0**	*1*
Second moment 13 × 13	0	0	0	1	**1**	0	0	0	0	**0**	0	0	0	0	**0**	*1*
Homogeniety 13 × 13	0	0	0	0	**0**	0	1	0	0	**1**	0	0	0	0	**0**	*1*
Mean3 × 3	0	0	1	1	**2**	0	0	0	0	**0**	0	0	0	0	**0**	*2*
Mean 5 × 5	0	0	0	0	**0**	0	0	0	0	**0**	0	0	0	1	**1**	*1*
Mean 7 × 7	0	1	0	0	**1**	0	0	0	0	**0**	0	0	0	0	**0**	*1*
Mean 9 × 9	0	0	0	0	**0**	0	1	0	0	**1**	0	0	0	0	**0**	*1*
Mean 11 × 11	0	0	0	0	**0**	0	0	0	1	**1**	0	0	0	0	**0**	*1*
Mean 13 × 13	0	0	0	0	**0**	0	0	0	0	**0**	0	0	0	1	**1**	*1*
Total per level per site	*0*	*1*	*1*	*4*	***6***	*0*	*2*	*1*	*2*	***5***	*0*	*0*	*0*	*2*	***2***	*13*
All indicators	**1**	**2**	**4**	**7**	**14**	**1**	**2**	**3**	**7**	**13**	**1**	**2**	**4**	**6**	**13**	**40**

slums, with most instances of incorrectly classified slum cells extending towards areas immediately surrounding these slums. There were almost no correctly classified instances in smaller slums. Similar to Nairobi, this may be in part due to the larger slums skewing the training data. Furthermore, large parts of slums may also contain vegetation, which could have led to much lower precision values. The relatively high precision value of 0.72 observed for Kisumu reflect the low number of type I errors for this site.

5.2. Discriminant analysis

As shown in Table 4, the classification accuracy of the discriminant analysis (DA) model was in the range of 0.22–0.36 and 0.66–0.79 for the precision and recall measures, respectively. Nairobi had the lowest classification accuracy followed by Kisumu and then Mombasa. These results are notably higher in comparison to the results obtained from the LR model. However, as Figure 3(b) shows, the DA model overfitted the data for all three study areas, producing a substantial number of type I errors, and which is also reflected in the low recall values. Many of these areas, as observed in high-resolution imagery, are built up with a mix of residential, commercial and industrial type land uses. Unlike the LR model, the DA model detected most of the smaller slums in all study locations. Figure 4 shows both slums and areas classified as slums overlaid onto a false color composite for Nairobi. Similar overfitting patterns were observed in the Mombasa and Kisumu sites. The high precision values reported for all three sites indicate that most correctly classified slum cells were relevant.

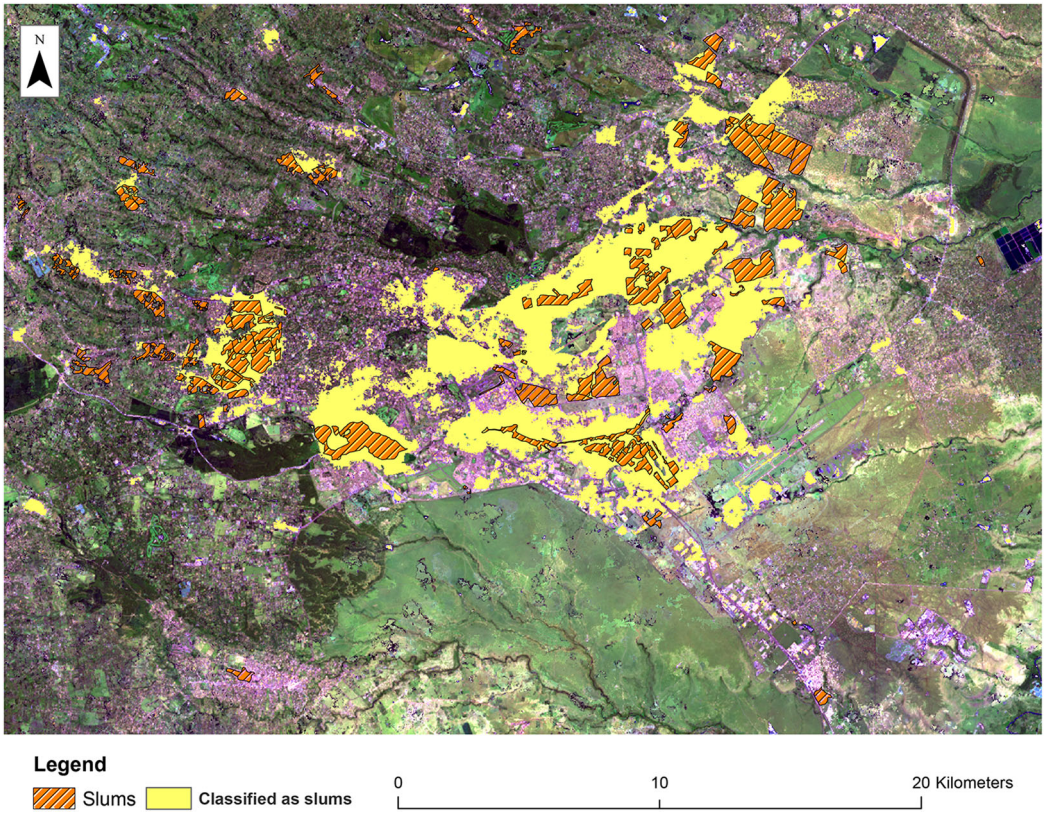

Figure 4. Urban and slum areas in Nairobi (False composite image created by stacking image bands 7, 6 and 4 from the Landsat 8 satellite. Pink to white tones highlight urban areas).

Table 5. Percent change in accuracy using the remote sensing data as the baseline.

RS + 1	RS + 2	RS + 3	All
6.55	6.81	6.85	6.52
1.55	2.38	2.51	2.49
1.62	2.05	2.087	2.13

Note: Number in the top row represent the addition of the top 1, 2 and 3 socio-economic indicators to the remote sensing data. Where 'RS' represents the remote sensing indicators and '+1', '+2' and '+3' represents the addition of the single, top 2 and top 3 socio-economic indicators, respectively.

5.3. See5

Overall, as Table 5 shows, the results of the See5 classification for all three sites were substantially higher than the results of the LR and DA models. All three sites had reported precision and recall accuracy values of at least 0.9. An examination of the spatial footprint of these results, shown in Figure 3(c), indicate that the See5 model was able to improve the overall mapping accuracy over the LR and DA models. Most type I and type II errors occurred along the boundaries between the slum and the non-slum areas.

To compare the results of the See5 decision tree between sites, the first four levels of each tree were extracted for each site (Figure 5). Table 5 shows a summary of the candidate indicators used at each site at each respective level. Such an approach has been found to be suitable for comparing trees, since the most important variables in the See5 tree appear towards the top levels of the tree, with the further movement towards the bottom leading to the more fine-tuned fitting of the data (Quinlan 2015). A comparison of all tree subsets in Table 5 shows that both socio-economic and remote sensing indicators were important for classifying a place as a slum. This table also shows that socio-economic indicators were more often used in comparison to remote sensing indicators for all sites. The total number of instances of socio-economic indicators were 2, 3 and 9 greater than the total number of instances of remote sensing indicators for the Nairobi, Mombasa and Kisumu

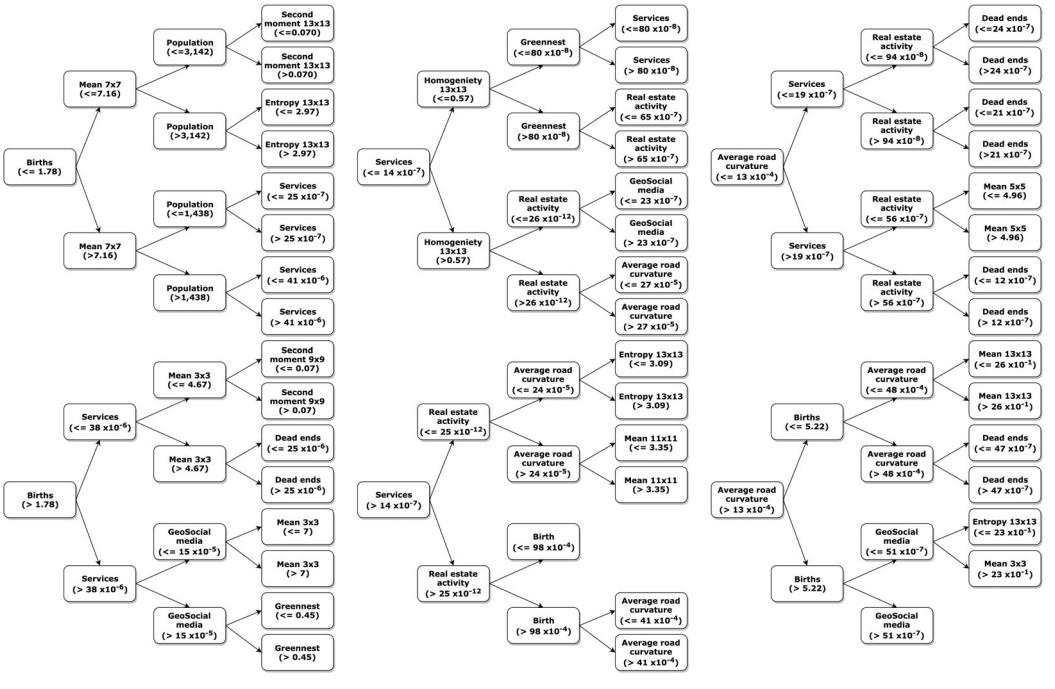

Figure 5. See5 tree up to 4th level for (a) Nairobi, (b) Mombasa and (c) Kisumu.

sites, respectively. Similarly, across all sites, the total number of instances of socio-economic indicators was more than double the total number of instances of remote sensing indicators (27 vs 13). Moreover, the most popular socio-economic indicator with respect to usage across all three sites was services (six instances), followed closely by dead ends (five instances) and average curvature of roads (five instances). The most popular remote sensing indicators across all sites were substantially lower: mean 3×3 texture and the greenest indicator, with two instances each.

Further examining each tree level, the root node (level 1), the most important indicator, for all sites was a socio-economic indicator. The root node for Nairobi was births, with larger values (refer to Figure 5) characteristic of larger slums. In slums in Nairobi, birth rates are much higher compared to other places in the city and other urban areas in Kenya. This is in part owing to a higher fertility rate of 3.5 for women in Nairobi slums compared to 2.6 elsewhere in Nairobi (African Population and Health Research Center 2014). For Mombasa, the root node was serviced, which may be partly due to slums in this city being more intermixed with the formal population compared to Nairobi. This collocation may have masked or even moderated some of the demographic patterns of slums there. In this case, variables collected over slums in Mombasa may have also contained information on adjoining formal populations. Finally, the root node for Kisumu was average road curvature. In Kisumu, there is a general lack of roads for this city, with most roads concentrated towards the downtown district area. Most slums in Kisumu are located in close proximity to this district and contain very irregular roads compared to the formal areas (United Nations 2005). This made higher average road curvature stand out as an indicator.

At level 2, besides Mombasa, socio-economic indicators were again used to help discriminate slums. In Nairobi, the use of the services indicator is consistent with the high density of places of worship, and private primary and health care facilities in many of the large slums (Adams, Islam, and Ahmed 2015; Bodewes 2013; Mombo 2012; Taffa and Chepngeno 2005; Tooley, Dixon, and Stanfield 2008). This is also supported by the mean 7×7 indicator, being reflective of roof cluster patterns in the larger slums. In Mombasa, as discussed above, some of the socio-economic characteristics may have been masked. However, larger homogeneity 13×13 and mean 9×9 indicators highlight the larger clusters of roof tops of dwellings in Mombasa compared to Nairobi. Finally, in Kisumu, most of the formal population live on the outskirts of the city and are more scattered (United Nations 2005). The higher birth rates for the large cluster of slums near the downtown district are therefore reasonable.

At level 3, the variables continue to discriminate slums mainly on their socio-economic characteristics. In Nairobi, population density within slums is much higher than their surroundings (Mundia and Aniya 2006). The 1 km $\times$ 1 km data used for extracting population density may have led this indicator to be pushed towards the lower branches of the tree. The higher geosocial media values, on the other hand, may be in part due to the much greater media coverage of slums in Nairobi, and which have been driven by humanitarian projects such as Map Kibera (2016). In other slums, this indicator may be less sensitive, since previous research has shown that the urban poor is often digitally left behind (Taubenböck et al. 2018). The use of the mean 3×3 indicator helps discriminate the smaller slums. At level 3, average curvature of roads continues to be important in discriminating slums in Mombasa, with a very limited presence of real estate activity. With the exception of population density, all previous level three indicators were used in Kisumu.

At level 4, both Nairobi and Kisumu emphasized more, albeit by a small difference, on the physical characteristics of slums. In the case of Nairobi, the limited vegetation and close proximity of dwellings stand out. In Mombasa, apart from geosocial media, a mixture of previous indicators resurfaced. In Kisumu, population density was the main driving factor. Overall, the results suggest that while each slum may be unique with respect to its own combination of physical and socio-economic properties, some amount of commonality between indicators, albeit at different levels of importance, do exist between slums at the different sites.

5.4. Indicator contribution

Following the application of See5 to the different study areas, the top three indicators with the most discrimination power for each site were retained. These indicators, as discussed in Section 5.3., were located towards the upper branches of the tree and had the highest number of correctly classified slum cases. Next, a See5 tree was built using the remote sensing indicators alone. Indicators were then added one at a time and a new decision tree built, with the most significant indicator added, followed by the next significant indicator and finally, the last indicator was added. In Nairobi, the order in which indicators were added were (1) Births, (2) Services and (3) Geosocial media. For Mombasa, this order was (1) Services, (2) Births and (3) Average road curvature. Finally, in Kisumu, the order of significant indicators was (1) Average road curvature, (2) Births and (3) Services.

Figure 6 shows the receiver operating characteristic (ROC) curves for each new decision tree built with the addition of the significant indicators. Table 6 shows the percent change in the area under the curve metric computed for each decision tree using the remote sensing data as the baseline. Considering the ROC curve for remote sensing alone as a baseline, Figure 6 shows how incorporating additional indicators can provide varying level of increase in the accuracy. Specifically, the largest improvement was observed with the addition of the most significant socio-economic indicator. For Nairobi, there was a 6% increase with the addition of this indicator while for Mombasa and Kisumu, this increase was only 2%. Accuracies further increased with the progressive addition of the next two significant socio-economic indicators, however, this increase was very small (<1%)

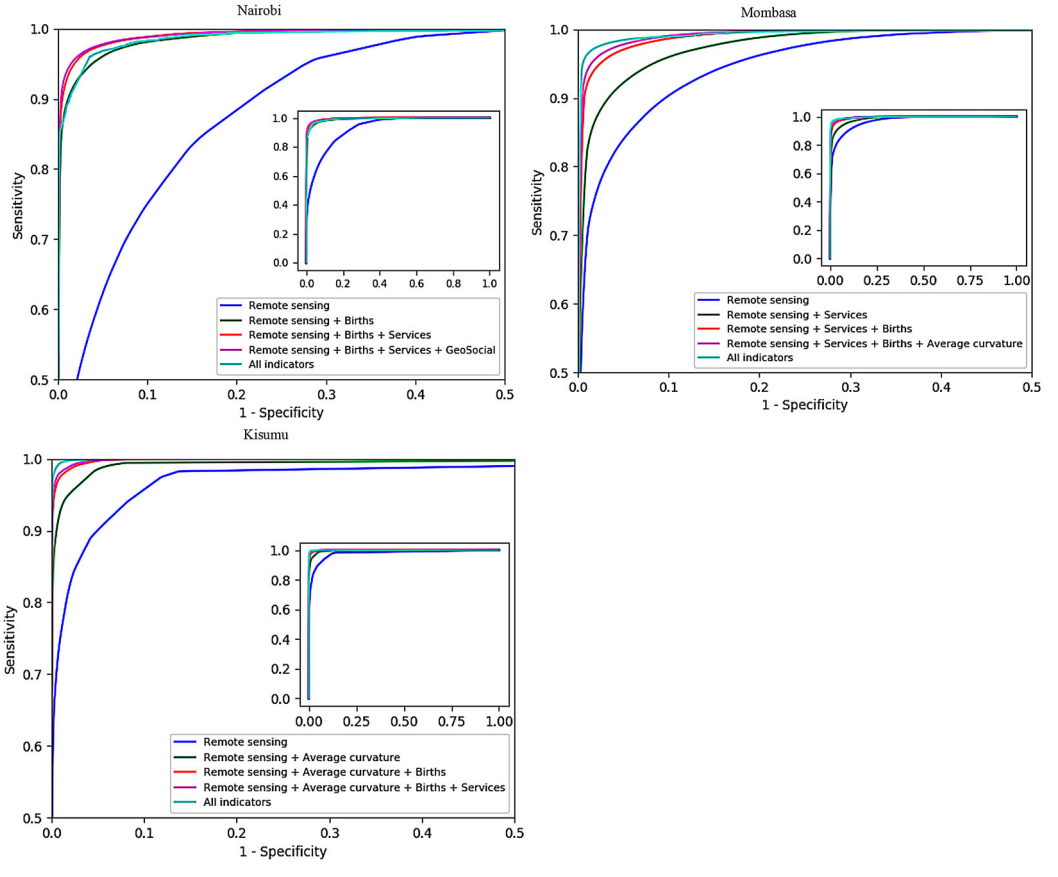

Figure 6. ROC curves for See5 trees built with the remote sensing indicators, the progressive addition of the most significant socio-economic indicators to the remote sensing indicators, and using all indicators for all three study sites. The inset image in each figure shows the full distribution of the data.

Table 6. Data mining models applied to other study areas.

	Nairobi		Mombasa		Kisumu	
Sites	Precision	Recall	Precision	Recall	Precision	Recall
Logistic regression						
Nairobi	NA	NA	0.00	0.07	0.00	0.00
Mombasa	0.01	0.19	NA	NA	0.00	0.01
Kisumu	0.03	0.18	0.22	0.29	NA	NA
Discriminant analysis						
Nairobi	NA	NA	0.34	0.30	0.07	0.06
Mombasa	0.8	0.14	NA	NA	0.46	0.30
Kisumu	0.17	0.18	0.73	0.22	NA	NA
See5						
Nairobi	NA	NA	0.25	0.34	0.03	0.42
Mombasa	0.37	0.12	NA	NA	0.08	0.16
Kisumu	0.46	0.14	0.46	0.14	NA	NA

Note: NA: not applicable.

as shown in Table 6. Furthermore, when all remote sensing and socio-economic indicators were used to classify slums, there was a slight decrease in accuracy for the Nairobi and Mombasa sites compared to the decision trees built using the remote sensing and top three socio-economic indicators. The Kisumu site, on the other hand, had only a very small improvement in accuracy when all indicators were used. Furthermore, in Nairobi, the accuracy of the See5 tree built using all indicators was comparable to the accuracy achieved with the use of remote sensing and the top three socio-economic indicators.

5.5. Model generalizability

Table 6 shows the classification results of models built using data at one study location and applied to the other two locations. This table shows that all models performed poorly in general. However, the DA model built using data from Mombasa and applied to the Nairobi site had a relatively high precision value of 0.80. A similar moderate precision value of 0.73 was obtained from the DA model built using data from Kisumu and applied to the Mombasa site. In the case of the former observed relationship, a more in-depth analysis of the standardized function coefficients for indicators for DA showed some highly positive and highly negative overlap values between the Mombasa and Nairobi sites. The application of the Nairobi model to Mombasa, however, produced poor classification results. This may have occurred due to the uniqueness of some of the socio-economic indicators included in the Nairobi model, which may have played a significant role in identifying slums in Nairobi, but not in Mombasa. With respect to the latter relationship between the DA results for Kisumu and Mombasa, it is uncertain why these results have occurred based upon a comparison of sites using variables' standardized function coefficients. Further exploration of the DA model for the Kisumu and Mombasa sites is therefore needed to better understand these results. However, it can generally be said that all study locations presented different geographies and with all slums having unique characteristics at each site, which is consistent with research on slums. Further, these results suggest that the models themselves were also specific to the data on which they were built.

6. Discussion and conclusion

With world population approaching eight billion people within the next decade, cities have become the focus of much attention regarding their ability to adequately provide for the increasing number of people. Cities in developing countries, in particular, are currently faced with various social, economic and developmental issues, which make it difficult to meet the needs of the growing population. This has resulted in the growth of large slum populations in less-developed countries. Consequently,

there is a pressing need for reliable information on slums to be used for monitoring and managing their growth.

To address this need for more reliable slum information, and to help overcome present data poverty issues in slum areas, this paper has demonstrated how open data can be transformed into candidate indicators and through data mining approaches, relevant indicators can be selected and combined for detecting and mapping slums. By using Kenya as our case study, we have shown that the approach presented here is applicable to less-developed countries where the issue of slums is most acute. Given the increasing availability of open sources of data in less-developed countries, the amount of data that is available for mapping and characterizing slums is also expected to increase. As such, the data mining approaches used in this paper could be used to help sieve through this large amount of data, providing a simple and objective means of selecting the most suitable indicators for mapping and monitoring slums based on their unique physical and socio-economic profiles.

Our results show that of the three data mining approaches that were evaluated, models derived using the See5 decision tree provided the highest classification accuracies for all three study sites. These accuracies were followed by the DA and LR models, respectively. Further, building on the strengths of the See5 model we showed that while slums within each study location had its own set of physical and socio-economic characteristics expressed in the subset of selected indicators that we evaluated, some of these characteristics were shared among slums at the different sites. For example, births as a socio-economic indicator were used within the first four levels of all decision trees built for each study site. However, this particular indicator occurred at different levels and with different thresholds within the See5 trees for each site. Thus, highlighting different levels of importance of this indicator for mapping slums at each site. These findings are important as they suggest that the present approach used in this paper can be used for laying the groundwork for creating an ontology for slums, thus complementing existing research on slums (e.g. Khelifa and Mimoun 2012; Kohli et al. 2012). Such work has mainly been reliant on the use of remote sensing imagery alone. Further, a comparison between remote sensing derived indicators and models derived from a combination of remote sensing and the most significant socio-economic indicators showed improved classification accuracies. The largest increase in accuracy was obtained when the most significant socio-economic indicator was added. These results highlight the need for including both physical and socio-economic indicators to inform a more comprehensive understanding and mapping of slums.

Various limitations have also been identified in this research, which also presents additional opportunities for future work. First, a 5-fold cross-validation approach was used to reduce issues with overfitting of the data. However, given the very skewed distribution of positive slum cases in this study, an optimum number of folds can only be determined with extensive experiments. Likewise, other approaches such as leave one out cross-validation, or penalization approaches such as Akaike Information Criterion or Bayesian Information Criterion may lead to different results.

Second, while the open data used in this research were available for all three study sites, additional data could be incorporated into a more detailed study (e.g. nighttime lights – Kuffer et al. (2018)). Depending on the subset of data used, this could lead to different overall accuracies and the most significant socio-economic indicators may also vary within the hierarchy of the See5 decision tree. Such data might also include the use of other texture measures not used in this study (e.g. local binary pattern – Leonita et al. (2018)), including, for example, those generated from other multispectral image bands.

A third limitation, and related to the previous point, is that some of the data used to derive indicators vary in spatial resolution, as much as 1 km × 1 km in some cases (e.g. population density). For some small slums or parts of slums, this coarse spatial resolution may have moderated their visibility in this data. There is, therefore, a need to further examine other higher resolution datasets, existing disaggregation methods (see e.g. Gaughan et al. 2014 for a review of population disaggregation methods), and the specific patterns that slums exhibit within their respective cities (Friesen et al. 2018) for detecting and mapping them. Such analysis should also explore the interplay between

the ability to detect slums, the spatial resolution of the data, and the a priori distribution of derived slum indicators and their sensitivity towards slums of different sizes within a given area (Wurm et al. 2017). This may provide additional insights as to why some indicators are better suited for mapping slums in one particular location compared to others. Finally, while this study centered on three large urban areas and their environs, the application of our methodology and derived indicators are expected to vary in rural settings, and in other cities with different physical and socio-economic profiles. However, as noted in Section 2, while the indicators may vary from one location to the next, our data-driven methodology is expected to be transferable.

Looking ahead, as more data capturing the physical and socio-economic properties of slums becomes available, a more in-depth decomposition into their unique fundamental building blocks can be derived and used for understanding each slum: for slums within the same geographic area (e.g. slums in Nairobi), different geographic areas within the same country (e.g. Nairobi vs Mombasa) and across different countries (e.g. slums in Kenya vs slums in India). Similar to work on the development of a global inventory on the spatial patterns and morphologies of slums derived from remote sensing imagery (e.g. Taubenböck, Kraff, and Wurm 2018), these profiles could then be used for mapping, characterizing, monitoring and comparing slums. Thus, providing important information for many urban initiatives such as the SDG and the WCCD. Such initiatives currently rely on data that may be out of date, inaccessible or unsuitable for mapping and monitoring slums across a global context. The approach used in this paper could potentially alleviate such issues, which are becoming increasingly prevalent in addressing the issue of slums and the sustainability of cities in general.

It is also important to note that while research on slums has been ongoing, most studies on slums only focus on a specific slum or slums within a specific city. Few have explored multiple slums in multiple cities (like in this paper), and even fewer have examined how slums change and evolve over time. The approach developed in this paper could be used as one part of a larger framework for understanding slums in this respect. For instance, candidate indicators used for characterizing slums could be studied over time and be used to evaluate the impact of policies or other interventions in slums. In so doing this paper lays the foundation for such an approach, by better understanding the specific characteristics that makes slums what they are, we can then address the specific social, economic, environmental and policy issues for slums in different cities, countries and regions of the world.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Ron Mahabir http://orcid.org/0000-0002-5553-5366
Anthony Stefanidis http://orcid.org/0000-0002-8165-0667
Arie Croitoru http://orcid.org/0000-0002-8470-9273
Andrew T. Crooks http://orcid.org/0000-0002-5034-6654

References

Adams, A. M., R. Islam, and T. Ahmed. 2015. "Who Serves the Urban Poor? A Geospatial and Descriptive Analysis of Health Services in Slum Settlements in Dhaka, Bangladesh." *Health Policy and Planning* 30 (Suppl 1): i32–i45. doi:10.1093/heapol/czu094.

African Population and Health Research Center. 2014. *Population and Health Dynamics in Nairobi's Informal Settlements: Report of the Nairobi Cross-sectional Slums Survey (NCSS) 2012 Report*. Nairobi: African Population and Health Research Center.

Afrobarometer. 2018. Accessed January 6, 2018. http://www.afrobarometer.org.

Ali, A. L., O. Hegazy, and M. N. Eldien. 2010. "Slum Prediction Using Integration Between GIS and ANN." Proceedings of the 7th international conference on Informatics and Systems, 1–8.

Angeles, G., P. Lance, J. Barden-O'Fallon, N. Islam, A. Q. M. Mahbub, and N. I. Nazem. 2009. "The 2005 Census and Mapping of Slums in Bangladesh: Design, Select Results and Application." *International Journal of Health Geographics* 8: 32.

Baud, I., N. Sridharan, and K. Pfeffer. 2008. "Mapping Urban Poverty for Local Governance in an Indian Mega-city: The Case of Delhi." *Urban Studies* 45 (7): 1385–1412. doi:10.1177/0042098008090679.

Bhaduri, B., E. Bright, P. Coleman, and M. L. Urban. 2007. "LandScan USA: A High-resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics." *GeoJournal* 69 (1–2): 103–117. doi:10.1007/s10708-007-9105-9.

Bhatta, B. 2009. "Modelling of Urban Growth Boundary Using Geoinformatics." *International Journal of Digital Earth* 2 (4): 359–381. doi:10.1080/17538940902971383.

Bodewes, C. 2013. *Civil Society in Africa: The Role of a Catholic Parish in a Kenyan Slum*. New Castle Upon Tyne: Cambridge Scholars Publishing.

Booth, C. 1903. *Life and Labour of the People in London*. London: Macmillan.

Chakraborty, A., B. Wilson, S. Sarraf, and A. Jana. 2015. "Open Data for Informal Settlements: Toward a User's Guide for Urban Managers and Planners." *Journal of Urban Management* 4 (2): 74–91. doi:10.1016/j.jum.2015.12.001.

Cohen, B. 2006. "Urbanization in Developing Countries: Current Trends, Future Projections, and Key Challenges for Sustainability." *Technology in Society* 28 (1–2): 63–80. doi:10.1016/j.techsoc.2005.10.005.

Cox, D. R. 1958. "The Regression Analysis of Binary Sequences." *Journal of the Royal Statistical Society* 20 (2): 215–242.

Craglia, M., L. Leontidou, G. Nuvolati, and J. Schweikart. 2004. "Towards the Development of Quality of Life Indicators in the 'Digital' City." *Environment and Planning B: Planning and Design* 31 (1): 51–64. doi:10.1068/b12918.

Crooks, A. T., A. Croitoru, A. Jenkins, R. Mahabir, P. Agouris, and A. Stefanidis. 2016. "User-generated Big Data and Urban Morphology." *Built Environment* 42 (3): 396–414. doi:10.2148/benv.42.3.396.

Crooks, A., D. Pfoser, A. Jenkins, A. Croitoru, A. Stefanidis, D. Smith, S. Karagiorgou, A. Efentakis, and G. Lamprianidis. 2015. "Crowdsourcing Urban Form and Function." *International Journal of Geographical Information Science* 29 (5): 720–741. doi:10.1080/13658816.2014.977905.

Dahmani, R., A. A. F. Fora, and A. Sbihi. 2014. "Extracting Slums from High-resolution Satellite Images." *International Journal of Engineering Research and Development* 10 (9): 1–10.

Du, S., F. Zhang, and X. Zhang. 2015. "Semantic Classification of Urban Buildings Combining VHR Image and GIS Data: An Improved Random Forest Approach." *ISPRS Journal of Photogrammetry and Remote Sensing* 105: 107–119. doi:10.1016/j.isprsjprs.2015.03.011.

Dubovyk, O., R. Sliuzas, and J. Flacke. 2011. "Spatio-temporal Modelling of Informal Settlement Development in Sancaktepe District, Istanbul, Turkey." *ISPRS Journal of Photogrammetry and Remote Sensing* 66 (2): 235–246. doi:10.1016/j.isprsjprs.2010.10.002.

Fayyad, U., G. Piatetsky-Shapiro, and P. Smyth. 1996. "From Data Mining to Knowledge Discovery in Databases." *AI Magazine* 17 (3): 37–54.

Fisher, R. A. 1936. "The Use of Multiple Measurements in Taxonomic Problems." *Annals of Eugenics* 7 (2): 179–188. doi:10.1111/j.1469-1809.1936.tb02137.x.

Flickr. 2017. Accessed February 15, 2017. https://www.flickr.com.

Frank, E., M. A. Hall, and I. H. Witten. 2016. *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*. Burlington, MA: Morgan Kaufmann.

Friesen, J., H. Taubenböck, M. Wurm, and P. F. Pelz. 2018. "The Similar Size of Slums." *Habitat International* 73: 79–88. doi:10.1016/j.habitatint.2018.02.002.

Gaughan, A. E., F. R. Stevens, C. Linard, N. N. Patel, and A. J. Tatem. 2014. "Exploring Nationally and Regionally Defined Models for Large Area Population Mapping." *International Journal of Digital Earth* 8 (12): 989–1006. doi:10.1080/17538947.2014.965761.

GFDRR. 2014. "Open Data for Resilience Initiative Field Guide: Global Facility for Disaster Reduction and Recovery." Accessed February 2, 2018. https://www.gfdrr.org/sites/gfdrr/files/publication/opendri_fg_web_20140629b_0.pdf.

Godin, B. 2003. "The Emergence of S&T Indicators: Why Did Governments Supplement Statistics with Indicators?" *Research Policy* 32 (4): 679–691. doi:10.1016/S0048-7333(02)00032-X.

Google Earth Engine. 2017. Accessed February 20, 2017. https://earthengine.google.com.

Google Map Maker. 2017. Accessed February 20, 2017. https://services.google.com/fb/forms/mapmakerdatadownload/.

Gotaway, C. A., and L. J. Young. 2002. "Combining Incompatible Spatial Data." *Journal of the American Statistical Association* 97 (458): 632–648. doi:10.1198/016214502760047140.

Government of Kenya. 2001. *Nairobi Situational Analysis: Consultative Report*. Nairobi: Government of Kenya.

Graesser, J., A. Cheriyadat, R. R. Vatsavai, V. Chandola, J. Long, and E. Bright. 2012. "Image Based Characterization of Formal and Informal Neighborhoods in an Urban Landscape." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5: 1164–1176. doi:10.1109/JSTARS.2012.2190383.

Grant, L. P., C. Gennings, and D. C. Wheeler. 2015. "Selecting Spatial Scale of Covariates in Regression Models of Environmental Exposures." *Cancer Informatics* 14 (Suppl. 2): 81–96.

Hacker, K. P., K. C. Seto, F. Costa, J. Corburn, M. G. Reis, A. I. Ko, and M. A. Diuk-Wasser. 2013. "Urban Slum Structure: Integrating Socioeconomic and Land Cover Data to Model Slum Evolution in Salvador, Brazil." *International Journal of Health Geographics* 12 (1): 45–45. doi:10.1186/1476-072X-12-45.

Haklay, M. 2010. "How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets." *Environment and Planning B: Planning and Design* 37 (4): 682–703. doi:10.1068/b35097.

Hassan, S., and R. Mahabir. 2018. "Urban Slums and Fertility Rate Differentials." *Population Review* 57 (2): 47–75. doi:10.1353/prv.2018.0006.

Hollander, J., M. Johnson, R. B. Drew, and J. Tu. 2017. "Changing Urban Form in a Shrinking City." *Environment and Planning B: Urban Analytics and City Science* 1–29. doi:10.1177/2399808317743971.

Isunju, J. B., K. Schwartz, M. A. Schouten, W. P. Johnson, and M. P. van Dijk. 2011. "Socio-economic Aspects of Improved Sanitation in Slums: A Review." *Public Health* 125 (6): 368–376. doi:10.1016/j.puhe.2011.03.008.

Jackson, S. P., W. Mullen, P. Agouris, A. Crooks, A. Croitoru, and A. Stefanidis. 2013. "Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information." *ISPRS International Journal of Geo-Information* 2 (2): 507–530. doi:10.3390/ijgi2020507.

Jenkins, A., A. Croitoru, A. T. Crooks, and A. Stefanidis. 2016. "Crowdsourcing a Collective Sense of Place." *PloS ONE* 11 (4): e0152932. doi:10.1371/journal.pone.0152932.

Kattan, J., R. Eid, H. R. Kourie, F. Farhat, M. Ghosn, C. Ghorra, and R. Tomb. 2016. "Mesotheliomas in Lebanon: Witnessing a Change in Epidemiology." *Asian Pacific Journal of Cancer Prevention* 17 (8): 4175–4179.

Khelifa, D., and M. Mimoun. 2012. "Object-based Image Analysis and Data Mining for Building Ontology of Informal Urban Settlements." *Proceeding of Image and Signal Processing for Remote Sensing* XVIII: 853711–853724.

Kitchin, R., T. P. Lauriault, and G. McArdle. 2015. "Knowing and Governing Cities Through Urban Indicators, City Benchmarking and Real-time Dashboards." *Regional Studies, Regional Science* 2 (1): 6–28. doi:10.1080/21681376.2014.983149.

KOD. 2018. "Kenya OpenData" Accessed May 21, 2018. https://opendata.go.ke.

Kohli, D., R. Sliuzas, N. Kerle, and A. Stein. 2012. "An Ontology of Slums for Image-based Classification." *Computers, Environment and Urban Systems* 36 (2): 154–163. doi:10.1016/j.compenvurbsys.2011.11.001.

Kotkin J. 2014. "Welcome to the Billion-man Slum." *New Geography*, August 25.

Kuffer, M., K. Pfeffer, and R. Sliuzas. 2016. "Slums from Space—15 Years of Slum Mapping Using Remote Sensing." *Remote Sensing* 8 (6): 455. doi:10.3390/rs8060455.

Kuffer, M., K. Pfeffer, R. Sliuzas, and I. Baud. 2016. "Extraction of Slum Areas from VHR Imagery Using GLCM Variance." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 9 (5): 1830–1840. doi:10.1109/JSTARS.2016.2538563.

Kuffer, M., K. Pfeffer, R. Sliuzas, I. Baud, and M. van Maarseveen. 2017. "Capturing the Diversity of Deprived Areas with Image-based Features: The Case of Mumbai." *Remote Sensing* 9 (4): 384. doi:10.3390/rs9040384.

Kuffer, M., K. Pfeffer, R. Sliuzas, H. Taubenböck, I. Baud, and M. van Maarseven. 2018. "Capturing the Urban Divide in Nighttime Light Images from the International Space Station." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 11: 2578–2586. doi:10.1109/JSTARS.2018.2828340.

Latin American Open Data Initiative. 2018. "Latin American Open Data Initiative." Accessed January 9, 2018. https://idatosabiertos.org.

Leonita, G., M. Kuffer, R. Sliuzas, and C. Persello. 2018. "Machine Learning-based Slum Mapping in Support of Slum Upgrading Programs: The Case of Bandung City, Indonesia." *Remote Sensing* 10 (10): 1522. doi:10.3390/rs10101522.

Logan, T. M., T. G. Williams, A. J. Nisbet, K. D. Liberman, C. T. Zuo, and S. D. Guikema. 2017. "Evaluating Urban Accessibility: Leveraging Open-source Data and Analytics to Overcome Existing Limitations." *Environment and Planning B: Urban Analytics and City Science* 1–17.

Lu, D., and Q. Weng. 2005. "Urban Classification Using Full Spectral Information of Landsat ETM+ Imagery in Marion County, Indiana." *Photogrammetric Engineering & Remote Sensing* 71 (11): 1275–1284. doi:10.14358/PERS.71.11.1275.

Mahabir, R., A. Croitoru, A. Crooks, P. Agouris, and A. Stefanidis. 2018a. "A Critical Review of High and Very High-resolution Remote Sensing Approaches for Detecting and Mapping Slums: Trends, Challenges and Emerging Opportunities." *Urban Science* 2 (1): 8. doi:10.3390/urbansci2010008.

Mahabir, R., A. Croitoru, A. T. Crooks, P. Agouris, and A. Stefanidis. 2018b. "News Coverage, Digital Activism, and Geographical Saliency: A Case Study of Refugee Camps and Volunteered Geographical Information." *PLoS ONE* 13 (11): e0206825. doi:10.1371/journal.pone.0206825.

Mahabir, R., A. Crooks, A. Croitoru, and P. Agouris. 2016. "The Study of Slums as Social and Physical Constructs: Challenges and Emerging Research Opportunities." *Regional Studies, Regional Science* 3 (1): 399–357. doi:10.1080/21681376.2016.1229130.

Mahabir, R., A. Stefanidis, A. Croitoru, A. T. Crooks, and P. Agouris. 2017. "Authoritative and Volunteered Geographical Information in a Developing Country: A Comparative Case Study of Road Datasets in Nairobi, Kenya." *ISPRS International Journal of Geo-Information* 6 (24): 24–25. doi:10.3390/ijgi6010024.

MajiData. 2016. Accessed July 9, 2016. http://www.majidata.go.ke.
Map Kibera. 2016. Accessed July 9, 2016. http://mapkibera.org/.
Mombo, E. M. 2012. "The Church and Poverty Alleviation in Africa." In *Anglican Women on Church and Mission*, edited by K. Pui-Ian, J. Berling, and J. P. Te Paa, Chap. 8, 134–149. New York: Morehouse Publishing.
Mooney, P., and P. Cocoran. 2014. "Has OpenStreetMap a Role in Digital Earth Applications?" *International Journal of Digital Earth* 7 (7): 534–553. doi:10.1080/17538947.2013.781688.
Mullen, W. F., S. P. Jackson, A. Croitoru, A. Crooks, A. Stefanidis, and P. Agouris. 2015. "Assessing the Impact of Demographic Characteristics on Spatial Error in Volunteered Geographic Information Features." *GeoJournal* 80 (4): 587–605. doi:10.1007/s10708-014-9564-8.
Mundia, C. N., and M. Aniya. 2006. "Dynamics of Landuse/Cover Changes and Degradation of Nairobi City, Kenya." *Land Degradation and Development* 17 (1): 97–108. doi:10.1002/ldr.702.
Nolan, L. B. 2015. "Slum Definitions in Urban India: Implications for the Measurement of Health Inequalities." *Population and Development Review* 41 (1): 59–84. doi:10.1111/j.1728-4457.2015.00026.x.
Olson, D. L., and D. Delen. 2008. *Advanced Data Mining Technique*. Heidelburg: Springer.
openAFRICA. 2018. Accessed January 12, 2018. https://africaopendata.org.
Open MENA. 2018. Accessed January 12, 2018. http://openmena.net/en/.
Openshaw, S. 1983. *The Modifiable Areal Unit Problem*. Vol. 38. Norwich: Geo Books.
OpenStreetMap. 2017. Accessed January 12, 2017. http://www.openstreetmap.org.
ORNL. 2017. Landscan. Accessed February 12, 2017. http://web.ornl.gov/sci/landscan/.
Owen, K. K., and D. W. Wong. 2013. "An Approach to Differentiate Informal Settlements Using Spectral, Texture, Geomorphology and Road Accessibility Metrics." *Applied Geography* 38: 107–118. doi:10.1016/j.apgeog.2012.11.016.
Parsons, T. 1997. "'Kibra Is Our Blood': The Sudanese Military Legacy in Nairobi's Kibera Location, 1902-1968." *The International Journal of African Historical Studies* 30 (1): 87–122. doi:10.2307/221547.
Patel, A., A. Crooks, and N. Koizumi. 2012. "Slumulation: An Agent-based Modeling Approach to Slum Formations." *Journal of Artificial Societies and Social Simulation* 15 (4): 2. doi:10.18564/jasss.2045.
Patel, A., N. Koizumi, and A. Crooks. 2014. "Measuring Slum Severity in Mumbai and Kolkata: A Household-based Approach." *Habitat International* 41: 300–306. doi:10.1016/j.habitatint.2013.09.002.
Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, et al. 2011. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research* 12: 2825–2830.
Pramanik, M. M. A., and D. Stathakis. 2016. "Forecasting Urban Sprawl in Dhaka City of Bangladesh." *Environment and Planning B: Planning and Design* 43 (4): 756–771. doi:10.1177/0265813515595406.
Pratomo, J., M. Kuffer, J. Martinez, and D. Kohli. 2017. "Coupling Uncertainties with Accuracy Assessment in Object-based Slum Detections, Case Study: Jakarta, Indonesia." *Remote Sensing* 9 (11): 1164. doi:10.3390/rs9111164.
Quinlan, J. R. 1996. "Improved use of Continuous Attributes in C4.5." *Journal of Artificial Intelligence Research* 4: 77–90. doi:10.1613/jair.279.
Quinlan, J. R. 2015. "Data Mining Tools See5 and C5.0." Accessed January 3, 2018. https://www.rulequest.com/see5-info.html.
R Core Team. 2018. Accessed January 10, 2018. https://www.r-project.org.
Ribeiro, B. M. G. 2015. "Mapping Informal Settlements Using WorldView-2 Imagery and C4.5 Decision Tree Classifier." Proceedings of the Urban Remote Sensing Joint Event, 1–4.
Robinson, W. S. 1950. "Ecological Correlations and the Behavior of Individuals." *American Sociological Review* 15 (3): 351–357. doi:10.2307/2087176.
Rupert, M. 2003. "Probability of Detecting Atrazine/Desethyl-Atrazine and Elevated Concentrations of Nitrate in Ground Water in Colorado." US Department of the Interior. Report no. 02-4269, Denver, CO.
Saraiva, C., and E. Marques. 2004. "A Dinâmica Social Das Favelas Da Região Metropolitana de São Paulo." Accessed May 27, 2017. http://neci.fflch.usp.br/sites/neci.fflch.usp.br/files/8306-20330-1-SM.pdf.
Schaffers, H., N. Komninos, M. Pallot, B. Trousse, M. Nilsson, and A. Oliveira. 2011. "Smart Cities and the Future Internet: Towards Cooperation Frameworks for Open Innovation." In *The Future Internet Assembly*, edited by J. Domingue, A. Galis, A. Gavras, T. Zahariadis, D. Lambert, F. Cleary, P. Daras, S. Krco, H. Müller, M. Li, H. Schaffers, V. Lotz, F. Alvarez, B. Stiller, S. Karnouskos, S. Avessta, and M. Nilsson, 431–446. Berlin: Springer.
Steele, J. E., P. R. Sundsøy, C. Pezzulo, V. A. Alegana, T. J. Bird, J. Blumenstock, J. Bjelland, et al. 2017. "Mapping Poverty Using Mobile Phone and Satellite Data." *Journal of the Royal Society Interface* 14 (127): 20160690. doi:10.1098/rsif.2016.0690.
Stevens, F. R., A. E. Gaughan, C. Linard, and A. J. Tatem. 2015. "Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-sensed and Ancillary Data." *PloS ONE* 10 (2): e0107042. doi:10.1371/journal.pone.0107042.
Taffa, N., and G. Chepngeno. 2005. "Determinants of Health Care Seeking for Childhood Illnesses in Nairobi Slums." *Tropical Medicine and International Health* 10 (3): 240–245. doi:10.1111/j.1365-3156.2004.01381.x.
Taubenböck, H., N. J. Kraff, and M. Wurm. 2018. "The Morphology of the Arrival City - A Global Categorization Based on Literature Surveys and Remotely Sensed Data." *Applied Geography* 92: 150–167. doi:10.1016/j.apgeog.2018.02.002.

Taubenböck, H., J. Staab, X. X. Zhu, C. Geiß, S. Dech, and M. Wurm. 2018. "Are the Poor Digitally Left Behind? Indications of Urban Divides Based on Remote Sensing and Twitter Data." *ISPRS International Journal of Geo-Information* 7: 304. doi:10.3390/ijgi7080304.

Taubenböck, H., and M. Wurm. 2015. "Globale Urbanisierung Markenzeichen des 21. Jahrhunderts." In *Globale Urbanisierung*, edited by H. Taubenböck, M. Wurm, T. Esch, and S. Dech, 5–10. Berlin: Springer Berlin Heidelberg.

Tooley, J., P. Dixon, and J. Stanfield. 2008. "Impact of Free Primary Education in Kenya. A Case Study of Private Schools in Kibera." *Educational Management Administration and Leadership* 36 (4): 449–469. doi:10.1177/1741143208095788.

UNDP. 2015. "Sustainable development goals." Accessed March 3, 2017. http://www.ua.undp.org/content/undp/en/home/mdgoverview/post-2015-development-agenda.html.

UNFPA. 2018. "About census. United Nations Population Fund – Myanmar." Accessed March 21, 2018. http://www.unfpa.org/transparency-portal/unfpa-myanmar.

UN Habitat. 2003. *The Challenge of Slums - Global Report on Human Settlements 2003*. London: Earthscan.

UN Habitat. 2006. *State of the World's Cities 2006/2007*. Nairobi: United Nations.

United Nations. 2005. *Situation Analysis of Informal Settlements in Kisumu*. Nairobi: United Nations.

United Nations. 2015. *The Millennium Development Goals Report 2015*. New York: United Nations.

UNSD. 2012. "Millenium Development Goal indicators." Accessed June 12, 2017. http://mdgs.un.org/unsd/mdg/Default.aspx.

Vatsavai, R. R. 2013. "Gaussian multiple instance learning approach for mapping the slums of the world using very high resolution imagery." Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1419–1426.

WCCD. 2017. "World Council on City Data." Accessed January 12, 2018. http://www.dataforcities.org.

Weeks, J. R., A. Hill, D. D. Ackerly, A. Getis, and D. Fugate. 2007. "Can we Spot a Neighborhood From the air? Defining Neighborhood Structure in Accra, Ghana." *GeoJournal* 69 (1–2): 9–22. doi:10.1007/s10708-007-9098-4.

WorldPop. 2017. "WorldPop Project." Accessed May 25, 2017. http://www.worldpop.org.uk.

Wu, X., V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, et al. 2008. "Top 10 Algorithms in Data Mining." *Knowledge and Information Systems* 14 (1): 1–37. doi:10.1007/s10115-007-0114-2.

Wurm, M., and H. Taubenböck. 2018. "Detecting Social Groups from Space – Assessment of Remote Sensing-Based Mapped Morphological Slums Using Income Data." *Remote Sensing Letters* 9 (1): 41–50. doi:10.1080/2150704X.2017.1384586.

Wurm, M., H. Taubenböck, M. Weigand, and A. Schmitt. 2017. "Slum Mapping in Polarimetric SAR Data Using Spatial Features." *Remote Sensing of Environment* 194 (1): 190–204. doi:10.1016/j.rse.2017.03.030.

Zetter, R., and F. A. De Souza. 2000. "Understanding Processes of Informal Housing: Appropriate Methodological Tools for a Sensitive Research Area." *International Planning Studies* 5 (2): 149–164. doi:10.1080/13563470050020167.