LINCOLN INSTITUTE
OF LAND POLICY

# Accuracy Assessment and Map Comparisons for Monitoring Urban Expansion: The Atlas of Urban Expansion and the Global Human Settlement Layer

Working Paper WP18AB1


Alejandro M. Blei
Marron Institute of Urban Management

Shlomo Angel
Marron Institute of Urban Management

Daniel L. Civco
Center for Land Use Education and Research

Yang Liu
Marron Institute of Urban Management

Xinyue Zhang
Marron Institute of Urban Management



**November 2018**

## Abstract

The availability of global high-resolution built-up area datasets provides researchers and policy makers a tool for monitoring progress on a number of sustainable development targets. These datasets are made possible by the application of increasingly sophisticated computer methods that eliminate the need for human intervention in classifying remotely sensed earth imagery. The European Commission's Global Human Settlement Layer (GHSL) is one such dataset with historical layers corresponding to the epochs of 2014, 2000, 1990, and 1975. We assess the accuracy of the GHSL landcover classification for the circa 2014 period, as well as the accuracy of the *Atlas of Urban Expansion* land cover classifications for the circa 2014 period, using reference map data that was manually digitized from high resolution satellite imagery in 200 global cities. We apply the urban extent methodology developed in *Atlas of Urban Expansion* to create GHSL urban extents and compare them to *Atlas of Urban Expansion* extents at the 1990, 2000, and 2014 time periods. The overall accuracies of the two datasets are essentially the same, never more than one percentage point apart. Urban extents created with GHSL data were smaller than *Atlas* extents in 2014, but larger in 2000 and 1990. Discrepancies between the datasets at the1990 period may require additional investigation to help establish the historical trend.



**Keywords**: Accuracy assessment, Landsat, GHSL, urban extent, urbanization, global comparison

## About the Authors

**Alejandro M. Blei** is a research scholar at the Marron Institute of Urban Management at New York University. He was a research coordinator for the Monitoring Global Urban Expansion research program, a tri-partite collaboration between New York University, UN-Habitat, and the Lincoln Institute of Land Policy. He is a co-author of the 2016 *Atlas of Urban Expansion.*
Address: 60 5th Avenue, 2nd Floor
New York, NY 10011
Telephone: 212-992-6872
Email: ablei@stern.nyu.edu

**Shlomo Angel** is a Professor of City Planning and the Director of the NYU Urban Expansion Program at the Marron Institute of Urban Management at New York University.
Email: sangel@stern.nyu.edu

**Daniel L. Civco** is a Professor Emeritus of Geomatics and former director of the Center for Land Use Education and Research (CLEAR) at the Department of Natural Resources and the Environment of the University of Connecticut.

Address: 1376 Storrs Road, U-4087
Storrs, CT 06269
Telephone: 860-486-0148
Email: daniel.civco@uconn.edu

**Yang Liu** is a research scholar and statistician at the Marron Institute of Urban Management at New York University.
Email: yl3371@nyu.edu

**Xinyue Zhang** is a Masters student at the NYU Center for Data Science and a graduate assistant at the Marron Institute of Urban Management at New York University.
Email: xz2139@nyu.edu

**Table of Contents**

**Accuracy Assessment and Map Comparisons for Monitoring Urban Expansion: The Atlas of Urban Expansion and the Global Human Settlements Layer**

## Introduction

Remotely sensed earth observation (EO) data drives our ability to systematically map and measure changes to the surface of the earth. The record of satellite based terrestrial observations extends backward nearly half a century, beginning with NASA's Landsat program and its Landsat 1 satellite, launched in 1972. Successive Landsat missions have provided a continuous stream of publicly available images with increasing spatial and spectral fidelity (Roy et al 2014). The analytical techniques to interpret the information collected from Landsat and related EO satellites are well developed and allow for the classification of image pixels into various land cover categories, including built-up areas. When mapped, the result can be used to assess the spatial extent of human settlements and its change over time.

While the classification of remotely sensed imagery can be used to identify amount and the location of built-up area, it does not tell us how this information relates to human settlements per se. It is the job of the analyst to interpret the data and make decisions about which areas to group together, which areas to leave separate, and why. These two tasks, the classification of remotely sensed imagery and the interpretation of these images to determine the outer boundaries of cities, and their change over time, represent the core work of the *Atlas of Urban Expansion—2016 Edition* (the *Atlas*).

The *Atlas* focused on a random stratified sample of 200 cities at three time periods, circa 1990, circa 2000 and circa 2014 with a view to making inferences about the universe of cities, or the set of all 4,231 cities in the world that were identified to contain populations of at least 100,000 in 2010 (Angel et al 2016a, Galarza et al 2018). Although the sample size was deemed sufficient for estimating global averages, our focus on 200 cities, a 4.7 percent sample, was largely a function of available resources. Our image classification procedures are labor intensive. It required approximately 25 hours to complete all classifications for a single city, from the initial step of image collection to the final step of post-classification manual editing. We studied cities circa the target date because we had to identify cloud free Landsat images over large study areas, which sometimes required going forward or backward a year or two from that target. Completing the *Atlas* also entailed a host of other tasks, including the collection of spatially explicit population data, the digitization of blocks and roads features from high resolution satellite imagery, and the completion of surveys of land and housing regulations and affordability across the 200-city sample, all of which placed additional constraints on the resources available to us.

The recent availability of global built-up area datasets with Landsat like spatial resolution (~30 meters), or better, where classification procedures have already been applied to the satellite images, represents a turning point for the study of settlements from EO data. Most importantly, it eliminates the costs of image classification to users. This means that that any city, group of cities, region, country, or conceivably the entire world, can be the focus of study where little else is required of the analyst except the extraction of data corresponding to the areas in question.

This has potentially profound implications for the type of work contained in the *Atlas* as it would allow for a substantial increase in the sample size, or the addition of country-based or region-based studies, either option at a minimal cost.

The key distinction between the classification methods applied in the *Atlas* versus those applied in the global datasets concerns the role of human decision-making in reviewing and adjusting outcomes generated by image analysis programs. The *Atlas* classifications are relatively time intensive because the unique spectral information within a city specific study area is analyzed on a case-by-case basis. Image analysis software initially clusters the data, but the analyst adjusts the software driven output to reflect human understanding of the built environment features in that particular location. It would not be possible for a human analyst to assess the entire surface of the earth in such a manner. The global maps we see today are made possible due to the application of increasingly sophisticated computer driven procedures to vast collections of remotely sensed images.

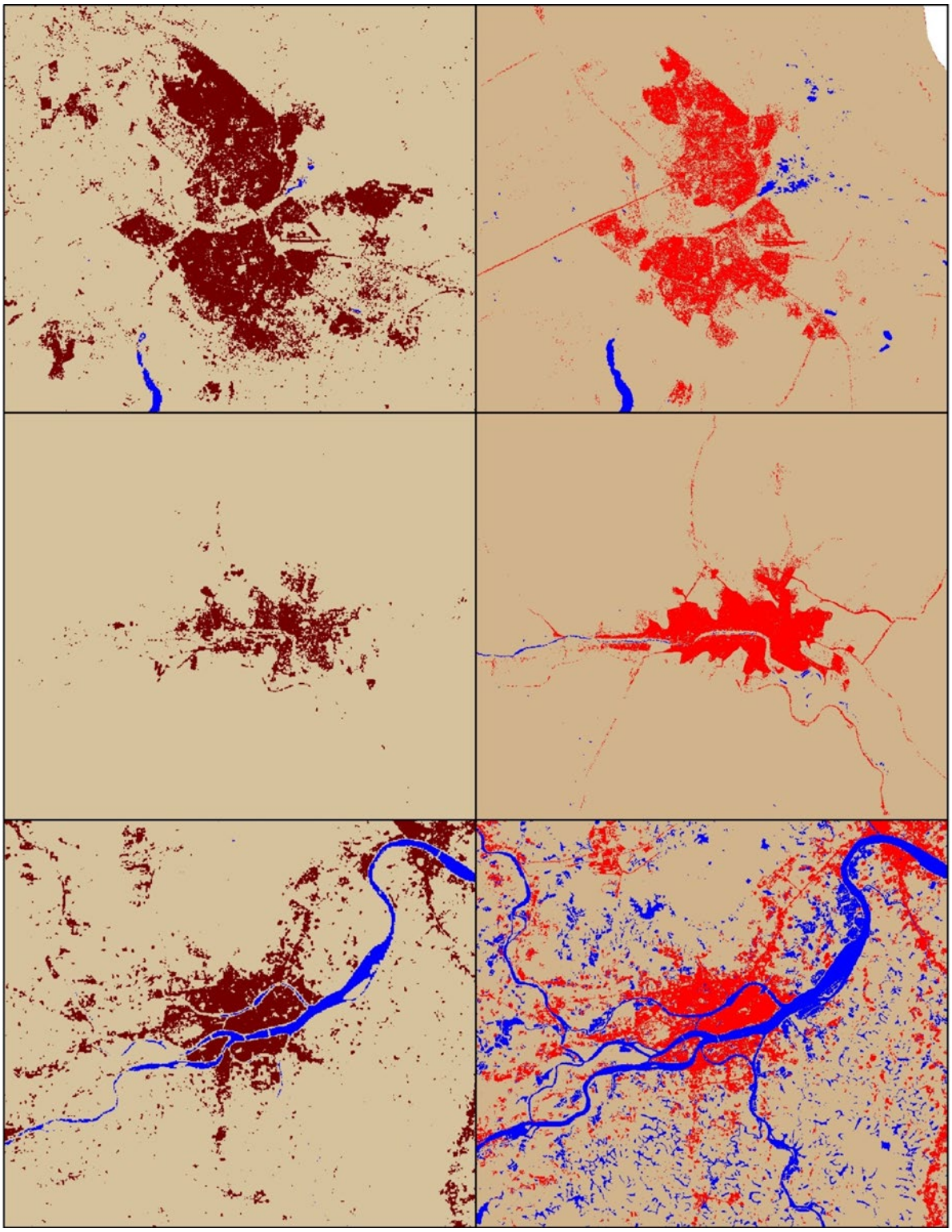
Before adopting a global built-up dataset in favor of our existing practices, we are curious to know how our map classifications compare and how differences in the map classifications may affect the size and the spatial agreement of cities' *urban extents*, a key feature of the *Atlas* methodology and a derived output of the land cover classifications. More concretely, in this paper we would like to address the following questions:

1. How accurate are the land cover classifications in the *Atlas* vs. land cover classifications from global built up maps?
2. How do the *Atlas* classifications and the global land cover classifications compare to each other?
3. When we apply the urban extent methodology to global-built up maps, how does the result compare to the findings obtained in the *Atlas*

We focus a single global urban map, the European Commission's Global Human Settlements Layer (GHSL), released to the public in 2016 (available at: http://ghsl.jrc.ec.europa.eu/datasets.php). The GHSL contains globally comprehensive 38-meter Landsat derived built-up area layers at four time periods: 1975, 1990, 2000, and 2014. The Landsat derived GHSL product provides the basis for the retrospective analysis of built up area worldwide and the 1990, 2000, and 2014 dates match our own study. Data for the 2016 GHSL built-up product, and for prospective products, is based, or will be based, on imagery collected from the European Space Agency's Sentinel satellite at a finer spatial resolution, of approximately 10-20 meters. We do not address the Sentinel derived built-up dataset in this paper. The procedures and added-value of Sentinel are discussed elsewhere (Pesaresi et al 2016; Corbane et al, 2017).

The GHSL is one of a number of products at increasingly finer spatial resolutions that map built-up area globally from EO data. Ten years ago, a survey identified ten such global maps, where

**Figure 1: GHSL 2014 (left) and *Atlas* (right) land cover classifications for Ndola, Zambia, Jun. 2014 (top row); Jequie, Brazil, Apr. 2014 (middle row); and Kaiping, China, Nov. 2014 (bottom row). Brown/red = built up, blue = water, and light brown = open space.**



the smallest spatial resolution among them was 309 meters and the typical map resolution was approximately 1 kilometer (Potere et al 2009). While the smaller resolution Landsat data had been available at that time, no group had processed the data to allow for it to be used as the basis of a global built-up map.

Today, the spatial resolution of global built-up maps has increased by a factor of 10 or more. Alongside GHSL there is GlobeLand30 (GL30), a Landsat derived 30-meter global land cover dataset produced by the National Geomatics Center of China, and Global Urban Footprint, a12-meter resolution dataset produced by the German Aerospace Center from the TerraSar-X and TanDEM-X satellites, which includes an urban areas layer based on images collected circa 2010. While GL30 contains information for ten land cover classes and GUF is at a fine spatial resolution, the GHSL is unique due to its historical data layers and its one-click ease of access to its entire dataset. This makes it an appropriate choice for studying built-up area change over decades at the global scale. Furthermore, GHSL plans to release new data layers on a yearly basis going forward while the frequency of the other products is unclear.

The accuracy of remotely sensed data, particularly in vast datasets that cover the entire earth, raises a number questions. It is relatively easier to calibrate detection algorithms when the analysis area is small, with relatively homogenous soils and vegetation, building materials, and geographic features. Increased heterogeneity in the input data, such as the combined area across multiple continents, entails a need for more complex algorithms that can discriminate between increased noise levels, which in turn make the target signal harder to detect. When Potere and Schneider (2007) compared the total amount of built up area in circa 2000 global maps, for example, where maps often shared the same inputs, they observed a range of approximately three million square kilometers between the lowest and highest estimates. Clearly, the different estimates cannot be simultaneously correct. Perhaps the differences between the newer generation of global built up maps are substantially smaller, though we do not know the answer with certainty, as such a comparison has not been undertaken, to our knowledge. A comparison of the GUF and 12-meter Sentinel derived GHSL suggests a high degree of correspondence between the two datasets, at least across urban and rural settings in Central Europe (Klotz et al 2016).

We had reason to believe that the Landsat classifications in the *Atlas* were of relatively high accuracy. The *Atlas* classification procedures were virtually unchanged from Angel et al (2005), whose authors obtained an average overall mapping accuracy of 89.2% from a pixel-based assessment in 12 cities. Later, Potere et al (2009) conducted an accuracy assessment on all 120 cities in Angel et al (2005), focusing on an area-based majority class assessment in place of pixel comparisons, obtaining an average overall accuracy of 87.1%. Members of the team that conducted the Angel et al (2005) classifications were involved in the classifications of Angel et al (2016), leading us to believe that the new Landsat classifications would be comparably accurate. When the project carried on longer than expected, a number of Landsat classifications were completed by the India Urban Expansion Observatory, a partner organization, and that work was conducted in the exact same manner.
An accuracy assessment and map comparison were not part of our original workplan but two factors influenced our decision to pursue the current exercise. The first was our discovery of the GHSL which occurred when the *Atlas* was in full production and nearly all classifications were

completed. Initial visual inspections of GHSL data for areas corresponding to a handful of *Atlas* cities suggested a high degree of correspondence between the two. We were intrigued by the prospect of a global time-series dataset with a similar spatial resolution that could be used to generate three-way classifications of built-up, water, and not built-up, in other words, a dataset that could provide the fundamental input for *Atlas* analyses. If we could prove to ourselves that the GHSL was of comparable accuracy or better accuracy than our own work, or that it met an acceptable level of accuracy, then we should be able to adopt the GHSL in the future with little hesitation. We were eager to obtain an answer to this question.

The second factor that influenced our decision to pursue this analysis was the recognition that the *Atlas* allowed for a novel and comprehensive method for carrying out such an exercise. Information collected for *Volume 2: Blocks and Roads*, could be used to assess the accuracy of all 200 *Atlas* classifications for the most recent time period. In each city, analysts had manually digitized and assigned land use categories to all block and road space within ten-hectare circular areas, called locales, distributed in a quasi-random fashion across the entire city. By aggregating the digitized features within an individual locale, it would be possible to differentiate the entire locale area into the binary categories of built-up and not built-up. The average city was assigned 87 locales and many cities contain more than 100. This means that the average city has at least 8.7 kilometers of manually digitized reference map data distributed in a quasi-random fashion across the city that could be used to assess the accuracy of the *Atlas* and GHSL land cover classifications. The broader analysis framework ensures that validation sites are approximately random distributed within cities and that the cities on which the analysis basis are distributed across world regions and across city population size categories.

We now have answers to the three questions posed earlier. First, although somewhat lower than other published results, the overall accuracy of the *Atlas* and GHSL is datasets is the essentially the same, never more than one percentage point apart. It should be noted that the similarity in overall accuracy masks differences in the accuracies of the built up and open space classes across the datasets. Second, the two datasets are quite similar to each other in overall terms, and largest difference between them concerns the identification of open space. Third, urban extents created with *Atlas* data were significantly larger than urban extents created with GHSL data in 2014, but *Atlas* extents were significantly smaller GHSL extents in 2000 and 1990.

The paper is structured as follows: section 2 provides an overview of the different datasets employed in the analysis; section 3 discusses the methods by which we conducted the accuracy assessment and map comparisons; section 4 describes the results obtained from different accuracy measures and map comparisons; section 5 interprets these findings, and section 6 concludes.

## Data

In this section we discuss the two levels of study sites—cities and locales, the *Atlas* and GHSL land cover data, and the reference map data digitized from high resolution satellite imagery. First, we describe how the global sample of cities was selected from the universe of cities and the transformation of Landsat data into urban clusters, and urban extents. Second, we describe how

intraurban study sites, or locales, were sampled from within cities' urban extents and how analysts digitized block and road features. Third, we discuss the GHSL dataset.

## The Global Sample of Cities

The rationale and procedures behind the creation of the global sample of cities were first described in Angel et al (2016a), "Chapter 2: The Global Sample of Cities, 1990 – 2014." A summary of the key points is presented below. Additional details concerning data acquisition and methods can be found in Blei et al (2018).

### The Universe of Cities

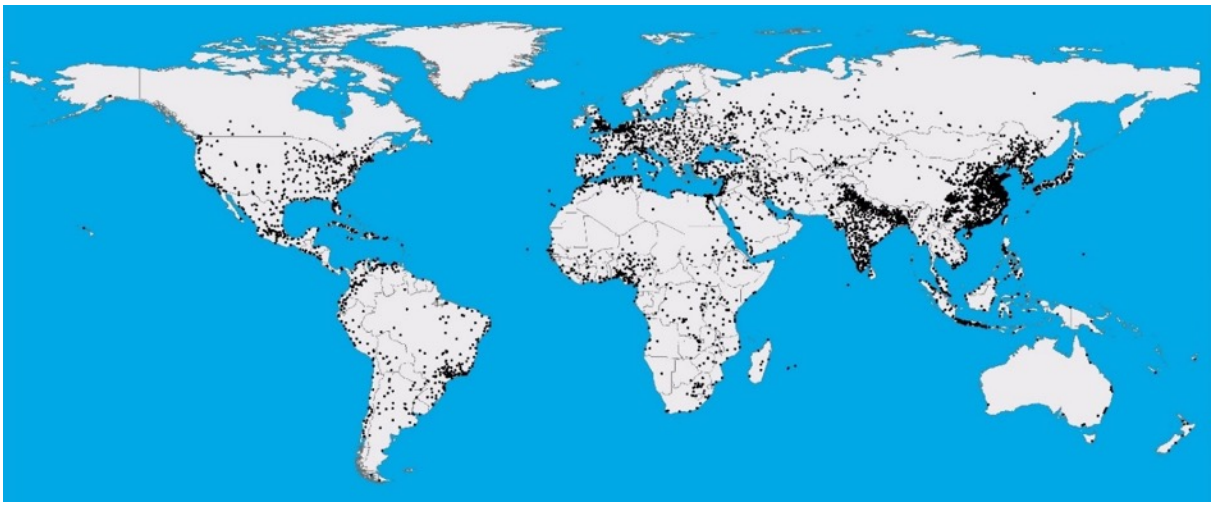
In our work, we have focused on cities as the unit of analysis. There is disagreement as to how many cities on earth there are because there is no universally accepted method for defining cities or identifying cities from satellite imagery. Governments and scholars have applied various combinations of the following criteria: population thresholds, administrative boundaries, density thresholds, commuting and activity patterns, and emerging concepts in urban studies (Parr 2007; OECD 2012; Uchida and Nelson 2008; Deuskar and Stewart 2016; and Taubenbock et al. 2014). We have chosen a definition that we could apply universally and consistently with existing data sources.

The 4,231 cities in the universe of cities represent contiguous or near-contiguous built-up areas of settlements that had populations of 100,000 or more in the year 2010. This area, or extent, is visible to the naked eye from high resolution satellite imagery, such as that which can be viewed on Google Earth or Bing Maps, and typically extends outward from a historical city center. A contiguous built-up area may include several municipalities and is neither constrained nor defined by administrative boundaries. Therefore, a single observation in the universe of cities may represent a number of adjacent municipalities.

To construct the universe of cities it was necessary to first identify candidate cities from lists of cities and towns, municipalities, metropolitan areas, and urban agglomerations with a reliable population estimate for 2010 or for which a population value at 2010 could be estimated. The three main data sources for this exercise were the UN Population Division, which provided information for settlements with populations of at least 300,000, the website www.citypopulation.de, which reproduces census data and census maps for all countries, and the Chinese Academy of Sciences which provided information for Chinese settlements.
Google Earth satellite imagery was used to inspect each candidate city, both to confirm its existence and to determine whether it should be merged with neighboring observations as part of a larger urban extent. Candidate cities below the population threshold that were not part of a larger extent were excluded from the analysis. In a small number of cases, those associated with cities that are part of larger metropolitan conurbations—such as the Northeast Corridor in the United States—the locally-defined metropolitan area boundary was used to differentiate one built-up extent from another, resulting in the separation of the New York and Philadelphia built-up areas, for example. Similar divisions were applied in China's Pearl River Delta region and in the Tokaido corridor in central Japan, as well as in a few other large conurbations where it was difficult to discern the boundaries of individual cities. In applying administrative or statistical boundaries in these limited cases, we acknowledge that a city's extent cannot extend endlessly; it

should roughly correspond to a commuting area or labor market area; in other words, the area linked together by social and economic spatial interaction.

**Figure 2: The universe of 4,231 cities with populations of 100,000 or more in 2010.**



Sampling Criteria

It was not possible to study each observation in the universe of cities and perhaps it should not be necessary, so long as there is a carefully constructed sample whose results can be generalized to the universe of cities as a whole. The universe of cities was organized along three strata with a view to selecting a representative sample.

The first stratum organized cities by eight world regions: (1) East Asia and the Pacific, (2) Southeast Asia, (3) South and Central Asia, (4) Western Asia and North Africa, (5) Sub-Saharan Africa, (6) Latin America and the Caribbean, (7) Europe and Japan, and (8) Land-Rich Developed Countries. Land-rich developed countries include the United States, Canada, Australia, and New Zealand. The regional categories roughly follow the divisions in the United Nation's *World Urbanization Prospects.* Cities were sampled from the eight regions in proportion to the population of the universe of cities in these regions.
The second stratum organized cities by city population size, of which there were four categories, roughly corresponding to small, medium, large, and very large: (1) 100,000 – 427,000; (2) 427,001 – 1,570,000; (3) 1,570,001 – 5,715,000; and (4) 5,715,001 and above. The total population of the universe of cities in each of these categories was approximately the same, about 622 million. An approximately equal number of cities was sampled from each of the four population size categories.

A third stratum was included in the sampling framework so that the sample would contain cities from countries with few cities as well as cities from countries with many cities. The number of cities in the country stratum contained three categories: (1) 1–9 cities; (2) 10–19 cities; and (3) 20 or more cities. Cities were sampled from these categories in proportion to the population of the universe of cities in these categories.

When combined, the eight regions, four population size groups, and three 'number of cities in the country' categories create 96 subcategories (8 × 4 × 3 = 96), or boxes, to which any observation in the universe of cities must belong. After distributing all 4,231 observations, 71 non-empty boxes remained. Two hundred cities were randomly drawn from these non-empty boxes in accordance with the sampling strategy.

**Figure 3: The global sample of cities.**

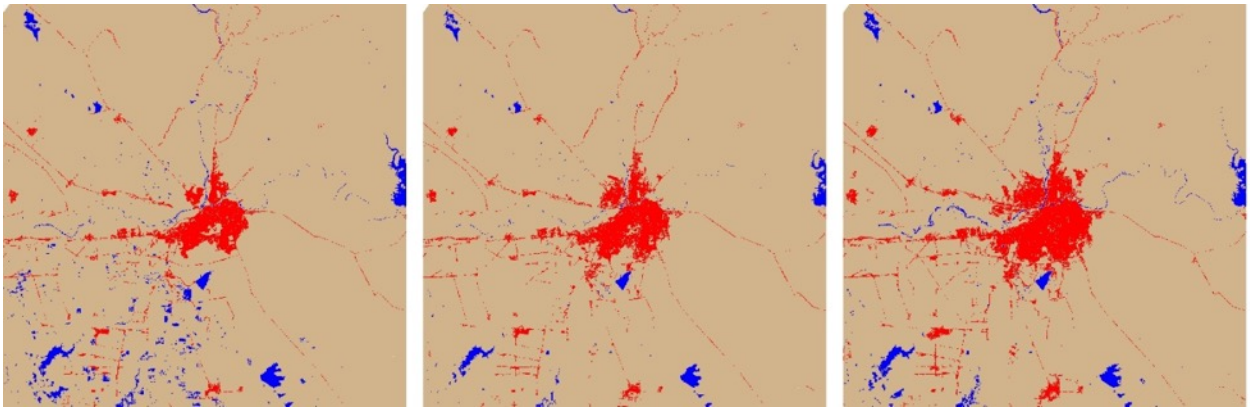
Landsat Data Collection and Classification

In *Vol 1: Areas and Densities*, we estimated urban extent populations by apportioning the population of spatially explicit population zones to all the built-up area within those zones and summing the apportioned values within the urban extent boundary. To streamline the process of satellite imagery collection and analysis, urban extent creation, and population apportionment, we first defined a city's study area by the set of population zones we believed would completely contain the urban extent. This decision was informed by an initial assessment of night lights data, which is known to overestimate built up area, and by verification of high resolution imagery. In a handful of cases, the final study area was determined through an iterative process. Upon creating the urban extent, we sometimes observed that it ran up against the study area boundary rather than terminating successfully within the study area. In these cases, we acquired additional spatially explicit population data to increase the size of the study area. Exceptions to this rule were those where the iterative process would have led to urban extents that contained more than one functional urban area or more than one metropolitan labor market. In these cases, we kept the study area boundaries fixed, using local definitions of metropolitan area boundaries or basing the decision on expert opinion.

Landsat scenes from Landsat 4, 5, 7, and 8 satellites, corresponding to dates circa 1990, 2000, and 2014 were downloaded from the United States Geological Survey's Earth Explorer website. Study area boundaries were superimposed on Landsat scenes corresponding to the three time periods. The intersected area, with an additional 1 km buffer, was selected for classification. Our objective was to extract three types of land cover categories from the Landsat images: water, built-up, and other/open space. Unsupervised classification techniques, where the analyst uses *a*

*posteriori* knowledge to label spectral classes generated by image analysis algorithms were performed using ERDAS Imagine software. The three-way classification of Culiacan, Mexico is shown in Figure 4.

**Figure 4: The three-way classification of Culiacan, Mexico in 1990 (left), 2000 (middle), and 2014 (right). Water = blue, open space = light brown, and built-up = red.**



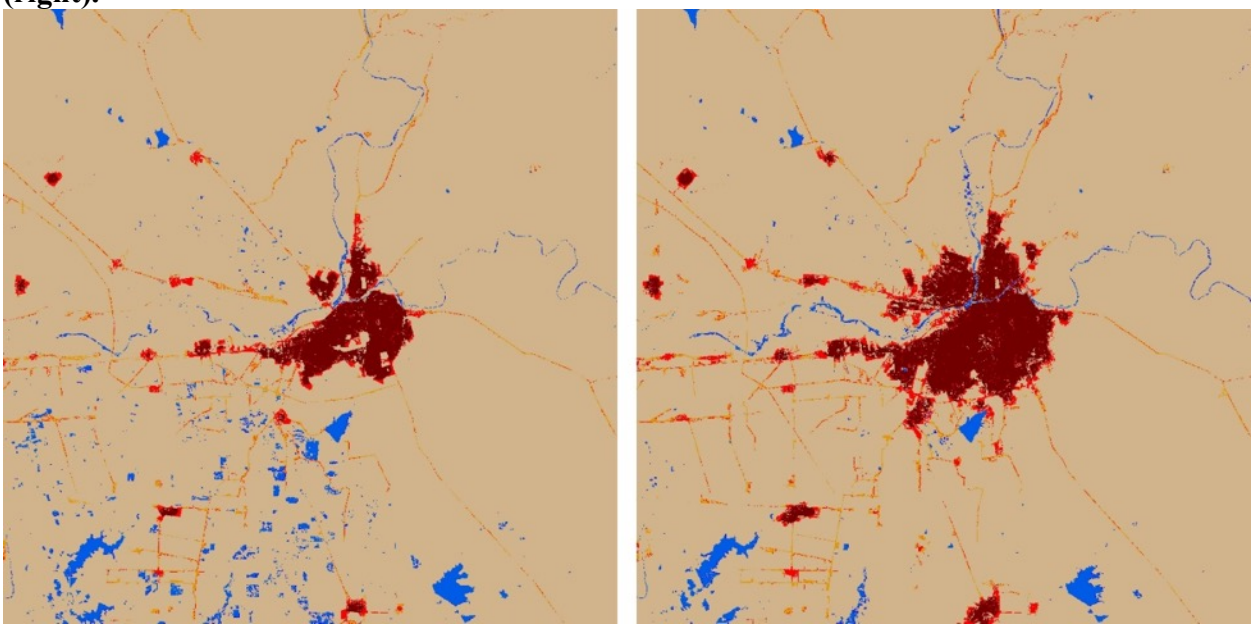
Landscape Analysis

The three-way classification of water, built-up, and open space was the input into a secondary analysis. This secondary analysis, or landscape analysis, sub-classified built-up and open space pixels into three categories each, allowing us to differentiate among different types of built-up and open space pixels. The sub-classification of the built-up class was based on the spatial density of built-up pixels within the Walking Distance Circle, defined as the 1 $km^2$ circle about an individual pixel. The three categories of built-up produced by the landscape analysis include:

1. *Urban* pixels, where the majority (> 50%) of pixels within the Walking Distance Circle are built up;
2. *Suburban* pixels, where 25-50% of pixels within the Walking Distance Circle are built-up; and
3. *Rural* pixels, where < 25% of pixels within the Walking Distance Circle are built-up.

The terms urban, suburban, and rural are used in the sense that the areas they generally correspond to our perceptions of what constitutes urban, suburban, and rural area in many cities throughout the world. The thresholds for the different categories are arbitrary and a different set of cutoffs would of course change the proportion of built up pixels in each category. We settled on these particular cutoffs after experimenting with different combinations of values in many cities, examining the output, and deciding which combination of values was associated with the most consistent and intuitive results. The three categories of built-up area pixels in Culiacan are shown in figure 5.

**Figure 5: The sub-classification of built-up area pixels into urban built-up (dark red), suburban built-up (red), and rural built-up (ochre) in Culiacan in 1990 (left) and 2014 (right).**

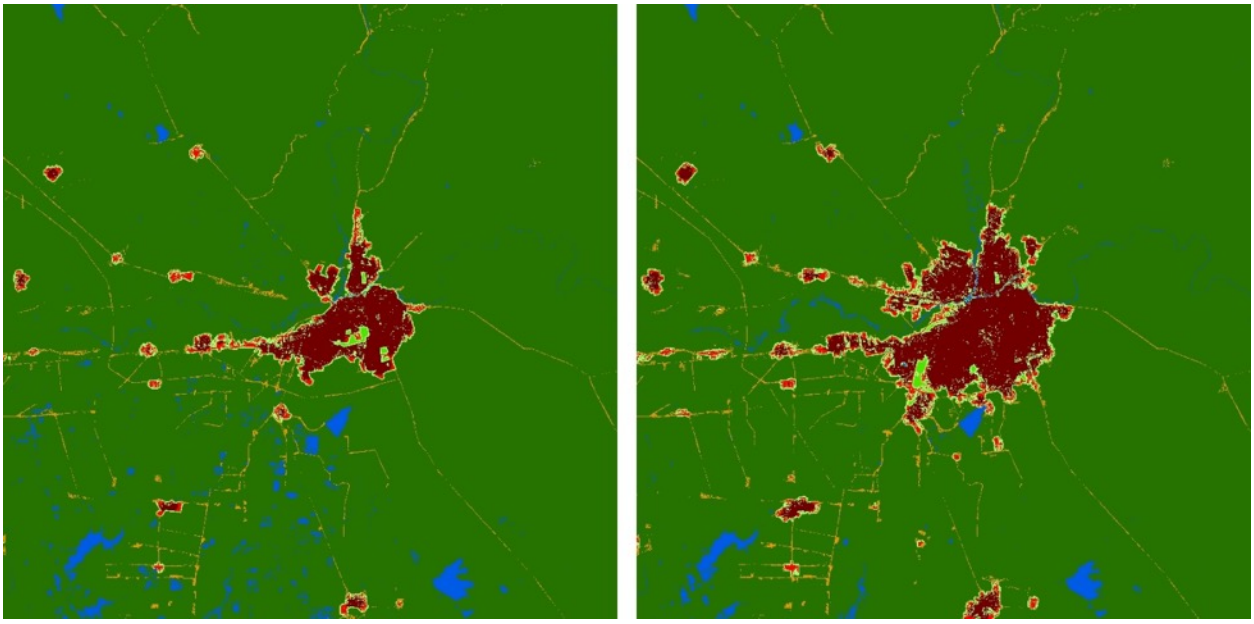


The three categories of open-space produced by the landscape analysis include:

1. *Fringe* open space pixels, all open space pixels within 100 meters of urban and suburban built-up pixels;
2. *Captured* open space pixels, clusters of open space pixels completely surrounded by fringe open space pixels less than 200 hectares in area; and
3. *Rural* open space pixels, all open space pixels that were neither fringe nor captured.

Taken together, the fringe and captured open space within a study area comprise the urbanized open space. Urbanized open space and rural open space together make up all of the open space within the study area. The three categories of built-up area pixels, the three categories of open space pixels, and water pixels are shown below in figure 6.

**Figure 6: The classification of open space into fringe open space (light green), captured open space (bright green), and rural open space (dark green) in Culiacan in 1990 (left) and 2014 (right).**



Urban Clusters and the Urban Extent Rule

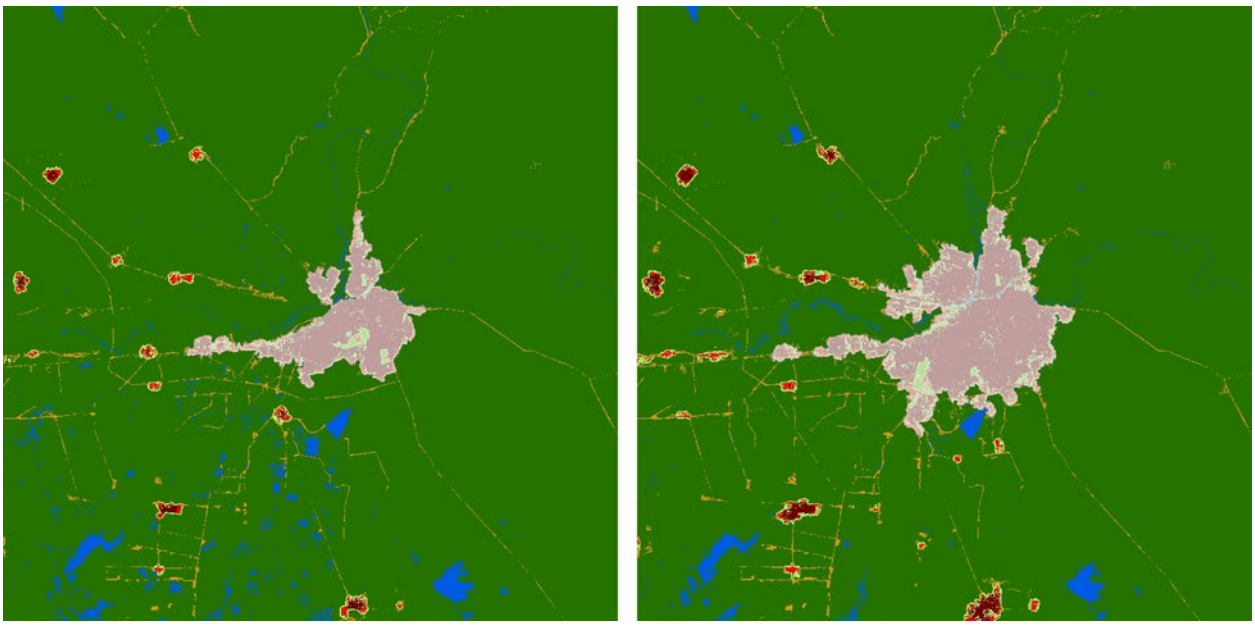
The landscape analysis differentiates built-up and open space pixels in a way that facilitates the creation of rules that can be used to identify clusters across the study area. We define urban clusters as discrete patches of urbanized open space, which by definition contain urban and suburban built-up pixels in their interior areas. There is no limit to the number of urban clusters within a study area; sometimes there is only one cluster and sometimes there are thousands. Figure 7 shows that there were several smaller clusters surrounding the main Culiacan cluster both in 1990 and 2014.

**Figure 7: Urban clusters across the Culiacan study area in 1990 (left) and 2014 (right).**

The urban extent represents a set of urban clusters within the study area. The challenge was to determine which clusters to include. We employed a rule based on the size and geographic proximity of clusters to each other to determine whether they should be grouped together into the same extent. We used this rule in the absence of globally available data that could be used to measure the strength of commuting ties between clusters, for example, or local knowledge about whether separate clusters should be considered to be one or two distinct cities.

More specifically, the decision of whether to group individual clusters together depended on an *inclusion rule.* We first generated a buffer around each cluster where the edge of the buffer area is always equidistant from edge of the cluster. The buffer distance for a given cluster is a function of the area of the cluster and it generates a buffered area equal to one-quarter the area of the cluster. The inclusion rule unites all clusters whose buffers intersect one another. The set of clusters with overlapping buffers forms an *urban extent*. The urban extent for the city in question is the grouping of clusters that contains the principal city's city hall location. Figure 8 shows the clusters that became the Culiacan urban extent in 1990 and 2014.

**Figure 8: The Culiacan urban extent in 1990 (left) and 2014 (right).**



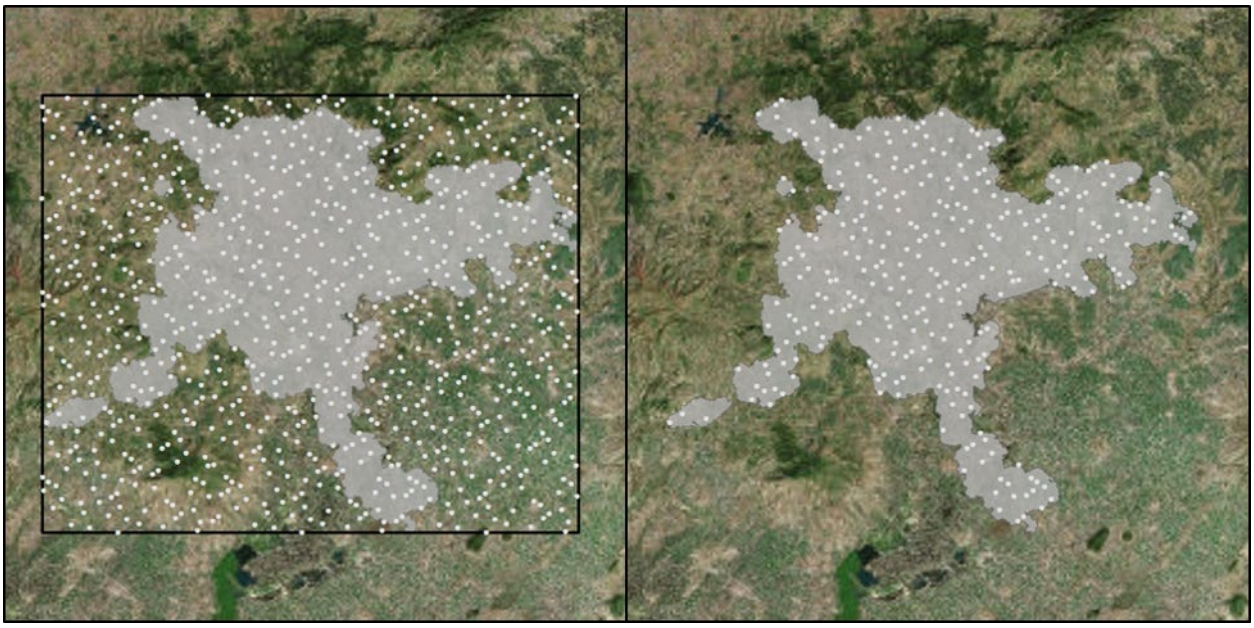
**Locales and the Intraurban Sampling Framework**

In the same way that we studied a sample of cities to draw inferences about the universe of cities, we studied a sample of intraurban locations to draw inferences about the spatial organization of blocks and roads across each city's urban extent. This sampling framework was developed for the *Atlas of Urban Expansion—Volume 2: Blocks and Roads.* Over each circular 10-hectare area sample site, or locale, analysts manually digitized blocks and roads features from high resolution satellite imagery. The spatial data for cities' blocks and roads features, as well as summary data tables of blocks and roads metrics may be downloaded from the *Atlas* website. For the present analysis, we use the information collected over these sampled areas as reference map data to assess the accuracy of the *Atlas* and GHSL land cover classifications. In this section, we discuss the generation of sample sites and the digitization procedures.

Bounding Box and Halton Sequence

The geographic coordinates of a bounding box that completely contains the urban extent was the basis of a Halton sequence of XY coordinate pairs within the box. The bounding box and the Halton points for Addis Ababa are shown on the left side of figure 9. We then focus on points that fall within the urban extent. Those points, shown on the right side of figure 9, represent the origins of potential sample sites within the urban extent. We employed a Halton sequence rather than randomly generated points for two key reasons. First, a Halton sequence produces a more even distribution of points across space compared to points generated by a truly random process, which results in some degree of clustering. Second, when the same initial XY coordinate pair is used to start the sequence, the points always occur in the same order and it is easy to maintain a relatively even spatial distribution of points by adding them in their sequential order.

**Figure 9: The Addis Ababa bounding box and its Halton sequence (left); Halton points within the Addis urban extent (right).**



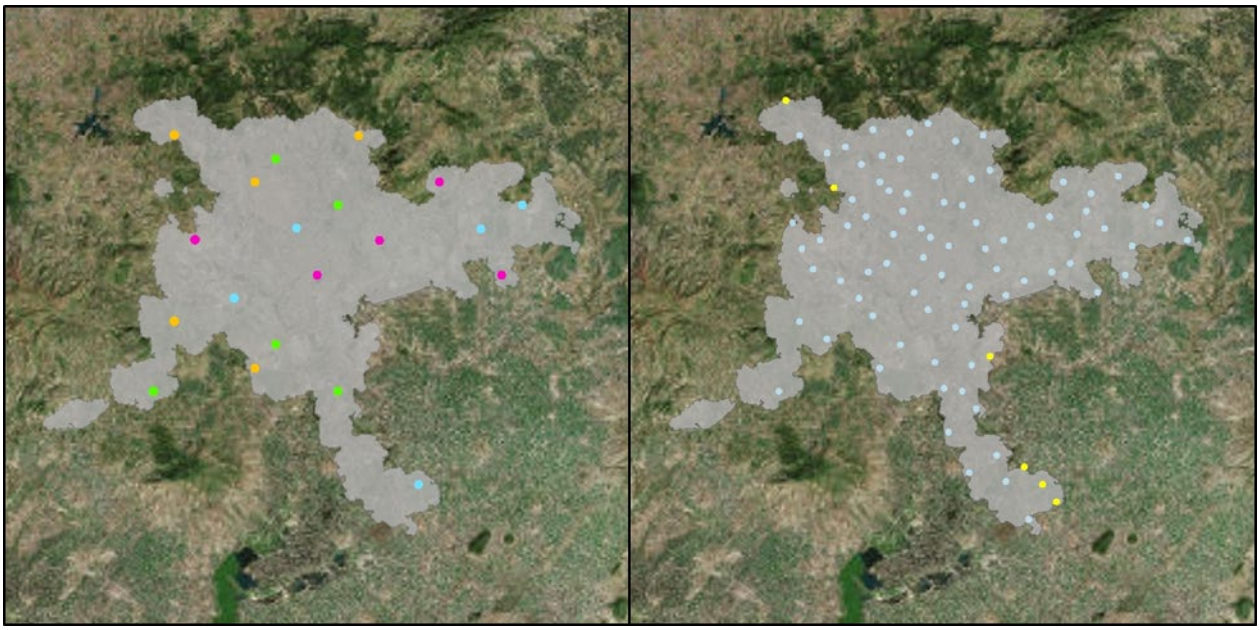
Locale Selection

Each Halton point within the urban extent was buffered by a radius of 178.4 meters to create the 10-hectare circular area called a *locale*. Analysts went through the ordered list of Halton generated locales and inspected the locale area against high resolution satellite imagery to determine whether the area was at least 80 percent built-up. Analysts performed this step because the original *Atlas* task was to obtain information about the spatial organization of blocks and roads. If the locale area was mostly open space, it would not contribute to our understanding of blocks and roads and it was skipped. Majority open space locales might fall within fringe or captured open space areas or in areas that were classified as built up but that corresponded to open space in actuality.

For the accuracy analysis, we identified these skipped locales and added them to the set of locales associated with a particular city. We added the skipped locales because we wanted the reference dataset to represent built up areas and open spaces relative to their actual distribution across the urban extent. In the original task, analysts selected at least 80 locales per city and digitized their blocks and roads to estimate various metrics. Smaller cities sometimes received fewer than that amount. After the initial allocation, locales were added based on available resources and *Blocks and Roads* metrics were adjusted accordingly.
The first twenty locales of the Halton sequence associated with the 2010 Addis Ababa urban extent are shown in figure 10 on the left. The first five points of the sequence are in green, the second five points in the sequence are blue, the third five points are in orange, and the fourth five points are in purple. This image illustrates how the sequential order of Halton points results in a relatively even distribution of locales across the analysis area. The image on the right shows the final distribution of 86 locales that were used for the Addis Ababa accuracy analysis. The six skipped locales added to the initial set of 80 are shown in yellow. The set of 86 locales used in

the Addis Ababa the accuracy analysis represent 8.6 $km^2$ of reference map data and 2.9 percent of the total area of the Addis Ababa urban extent of 296 $km^2$.

**Figure 10: The first 20 locales based on the Addis Ababa Halton sequence (left) and the final set of 86 Addis locales.**



Across 194 cities that were included in the accuracy analysis, the average city contained 87 locales and 11 of those locales, or 13 percent, were added because they had been skipped during the initial locale selection. The locale totals may appear somewhat lower than expected based on the initial and additional allocation targets. This is explained by data programming idiosyncrasies unique to the accuracy analysis and the need to eliminate a number of locales that did not meet specified criteria.

Locale Digitization and Labeling

The selection of locales and the digitization of their interior space was carried out by a group of analysts who were trained by the project team and instructed to follow a set of guidelines contained in an analyst manual. Digitization was carried out in the Java OpenStreetMap (JOSM) program, an editing tool for OpenStreetMap, a computer mapping application that uses Bing Maps as its satellite imagery layer. The initial rationale for the digitization of block and roads was to obtain information to estimate various metrics contained in *Vol 2: Blocks and Roads.* We have adopted that information and reorganized it for the accuracy analysis.

Analysts were instructed to first segment the locale area into street space and block space. Street space was taken to include all area conforming to the right-of-way, meaning any area that is currently used or could be potentially used by vehicles or pedestrians for travel. This included roadways, continuous sidewalks, bike paths, and street parking lanes. Paths within parks or spontaneous paths within unbuilt blocks were not included.

Block space, or the areas bounded by street space, was differentiated into open space and built up area. Open space refers to unbuilt areas and may represent open countryside, forests, crops, parks, cleared land, and water bodies. Built-up area was differentiated along two broad classes: non-residential and residential. Non-residential refers to built-up areas whose purpose is not residential and includes industrial parks, airports, sports facilities, malls and plazas, shopping centers, parking lots, and playgrounds, among others. Unlike the non-residential category, residential areas were sub-categorized into different classes depending on the form of the structures, the relationship of the structures to each other, the homogeneity of the structures, and plot sizes. These four categories included: atomistic, informal subdivision, formal subdivision, and housing project and are further described in *Vol 2: Blocks and Roads*, Chapter 3: "Understanding and Measuring Urban Layouts". The residential and non-residential categories comprise all non-street built-up area within the locale. All street space was treated as built-up. Therefore, the sum of street space, non-residential, and residential represents all built up area within the locale. The sum of built up area and open space represents the entire locale area. In this analysis, we are interested in the binary classification of locale space into built up area and open space. The aggregation of subcategories of built up area into a single built up class and the binary distinction of built up and open space in a single locale is shown in figure 11.

**Figure 11: An Addis Ababa locale boundary (left); the digitization of its block space into residential (pink), non-residential (yellow), street space (hollow), and open space (green), (middle); and the binary built-up area/open space classes (right).**

Reference Data Imagery Date

An important question for the accuracy analysis concerned the imagery date on which the digitization of locale features was based. When the temporal distance between the comparison maps dates (*Atlas* and GHSL classification) and the reference imagery increases, the higher the likelihood, all things being equal, that the accuracy assessment yields errors. Intuitively, the direction and distance matter for interpreting the type of error. If the reference imagery date precedes the map classification date, then perhaps the classification identifies built up area but the reference map does not. This ordering of dates may therefore be associated with a higher incidence of built-up area commission errors. Of course it is possible that the area was built-up in the intervening period, but there is no way to know for certain whether the error represents true error or a false error. All we know is that we might expect a higher incidence of commission

errors than compared to the case when the classification date precedes the ground truth date. In that case, the classification may identify no built-up area while the ground truth does. Of course it is possible that nothing existed at the time of the map classification and that the area was built in the intervening period. We might expect more errors of built-up area omission with this ordering of dates.

We could obtain the dates of the Bing satellite images used for digitization using the Bing Areal Imagery Analyzer for OpenStreetMap, a web-based resource. A computer program scraped the imagery dates associated with locale centroids from the Bing Areal Imagery Analyzer and appended this information to locales. This information was extracted after the *Atlas* was completed, during the last half of 2017 and the first months of 2018. All digitizations for the *Atlas* were completed by August 2016. This means that if a city is associated with a ground truth imagery date after August 2016, we know that it cannot truthfully represent the actual imagery used. Approximately 10 percent of cities' imagery dates fall into this range.

Table 1 summarizes information about reference imagery dates across and within cities, as well the temporal relationships among cities' reference imagery dates, NYU classification dates, and GHSL classification dates. To our surprise, there can be a surprising amount of variation in the reference imagery dates within a city. The first column summarizes the average reference imagery date across cities. For all cities, the average reference imagery date is August/September 2013. This is very similar to the average NYU date classification date of September 27, 2013.

**Table 1: Summary statistics for the variation in ground truth dates across and within cities and differences between ground truth dates and classification dates.**

| | Avg. ref. map date across cities | Ref. map temporal range (yrs), within cities | NYU classification date | GHSL classification date | \|Ref. map - NYU\| (yrs) | \|Ref. map - GHSL\| (yrs) | \|NYU - GHSL\| (yrs) |
|---|---|---|---|---|---|---|---|
| Avgerage | 8/31/2013 | 1.8 | 9/27/2013 | 7/1/2014 | 2.0 | 1.8 | 0.9 |
| Min | 12/31/2003 | 0.0 | 1/1/2009 | 7/1/2014 | 0.0 | 0.0 | 0.0 |
| 10th prctle | 11/11/2010 | 0.0 | 12/25/2011 | 7/1/2014 | 0.3 | 0.3 | 0.2 |
| 25th prctle | 2/9/2012 | 0.0 | 8/1/2013 | 7/1/2014 | 0.8 | 0.6 | 0.2 |
| Median | 11/24/2013 | 1.1 | 1/1/2014 | 7/1/2014 | 1.7 | 1.7 | 0.6 |
| 75th prctle | 3/22/2015 | 2.9 | 5/1/2014 | 7/1/2014 | 2.6 | 2.6 | 1.0 |
| 90th prctle | 8/21/2016 | 5.5 | 10/1/2014 | 7/1/2014 | 4.1 | 3.6 | 2.8 |
| Max | 7/12/2017 | 9.8 | 6/1/2016 | 7/1/2014 | 10.1 | 10.5 | 5.5 |

Since we do not know the actual dates associated with GHSL classifications we assigned them a date corresponding to the midpoint of 2014. In Pematangsiantar, Indonesia, the average reference imagery date was observed to be the earliest, from late 2003, while in Sao Paulo it was observed to be the latest, form mid-2017. Within cities, the average variation of imagery dates is 1.8 years and the median variation is 1.1 years. In 19 cities it was greater than 5.5 years. At the city level, the median pairwise difference between the reference imagery date and the classification date was 1.7 years for both the *Atlas* and GHSL data. The difference is even smaller, 0.6 years, when we compare the *Atlas* and GHSL data directly. While there a few observations where the reference imagery date is relatively far away from the classification date, this difference is less

than 2.6 years in 75 percent of cities. We believe this represents good temporal agreement among the three datasets and that it should minimize the effect of temporal variation in explaining errors.

**GHSL Dataset**

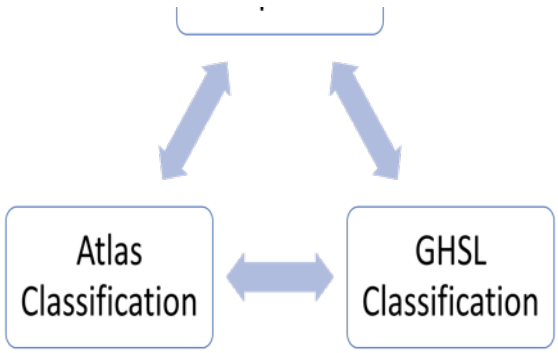
The Global Human Settlement Layer (GHSL), produced by the Joint Research Center of the European Commission, is a project aimed at monitoring human presence on the planet over time. It is comprised of gridded layers of built-up area and population, at resolutions of 38.2 meters and 250 meters respectively, which feed into a 1-kilometer gridded settlement classification model (Pesaresi et al 2016a). The built-up area layer is based on images collected from Landsat satellites over a period of more than 40 years. The output of the analyzed images has been been grouped into four collections, corresponding to the epochs of 1975, 1990, 2000, and 2014. The actual dates of images used to create each collection may vary forward or backward from the target by a number of years.

The GHSL identifies built-up area from the satellite images using the approach of symbolic machine learning which was designed for remote sensing big data analytics (Pesaresi et al. 2016b). The association analysis techniques employed are commonly used in bio-informatics for uncovering relationships between environmental effects and gene expression. The GHSL supervised detection methods are automatic and have been calibrated with a diverse set of fine scale and broad scale training data, including: OpenStreetMap data for roads, settlement places, and urban cover; settlement locations from Geonames; settlement polygons from diverse sources; MODIS urban extent data, MERIS Globcover, and Landscan population density grids. In this study we focus exclusively on the 38.2 meter built up grid with data for 1990, 2000, and 2014, which was downloaded from the GHSL online data portal.

## Method

In this section we discuss how we performed (1) the accuracy assessment of the *Atlas* and GHSL classifications and (2) intermap comparisons of *Atlas* and GHSL data, for both the land cover classifications and a derived outputs of the classifications, or the urban extents. We sough to compare each data set to all others, as depicted in figure 12.

**Figure 12: The relationships explored among the three datasets.**



**Accuracy Assessment**

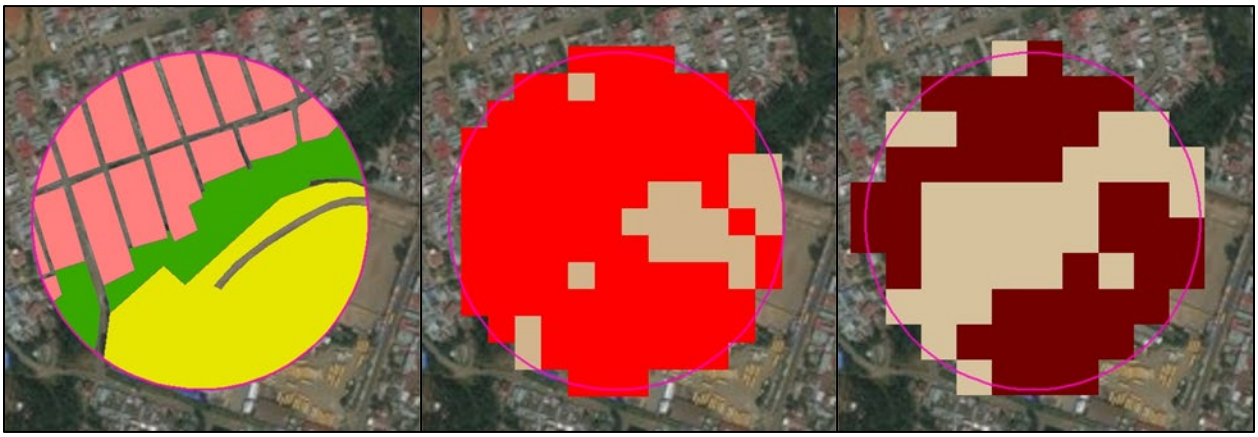
Map accuracy is typically assessed by comparing a produced map against a reference map or ground truth data. The task is more complicated when the analysis locations are globally distributed than when the locations can be field visited, observed, and recorded in person. High resolution satellite imagery helps us address this problem, though perhaps not definitively. Failure to account for the accuracy of a map classification leaves in doubt the confidence in the result and the generalizability of the conclusion.

Even when reference data are available, assessing map accuracy is not always straightforward. To be sure, there are several factors to consider in an accuracy assessment, including ground data collection, classification scheme, spatial autocorrelation, sample size, and sampling scheme (Congalton, 1991). With regard to mapping human settlements, however, how should map accuracy be assessed? Should we care about the accuracy of smallest possible map unit, that is to say, whether an individual pixel in the produced map is accurate with respect to the reference data or ground truth? Or perhaps there should be greater focus on whether the amount of built-up area over some area interest matches the amount of built-up area indicated by a reference map? There are benefits to both approaches to understanding map accuracy. On the one hand, the accuracy of individual pixels matters as they are the fundamental mapping unit and the basic input into our landscape analysis, which in turn influences the generation of urban extents. On the other hand, we are mostly interested in map accuracy in the sense that we want a reliable estimate of urban extent area, so that we can measure its change, and population density change, over time. From this perspective, a broader measure of area agreement would be more important than agreement at a small spatial scale such as individual pixels.

Accuracy comparisons can focus on the exact spatial allocation of a particular class or on broader measures of quantity agreement of classes (Pontius and Millones 2011). These comparisons may occur at varying level spatial scales. In other words, at the pixel level there may disagreement, but within a larger area such as a locale, there may be 100 percent quantity agreement between pixel classes despite spatial disagreement between pixel classes. Indeed, it would be possible to have comparisons with 100 percent spatial disagreement and 100 percent quantity agreement. Such mismatches are less problematic at smaller spatial scales but harder to reconcile as the spatial scale of analysis increases. In this accuracy assessment we focus both spatial allocation and quantity comparisons within locales.

Conducting the accuracy assessment presented a fundamental problem as the three datasets in question: the reference map data, the *Atlas* classifications and the GHSL classifications are at different geometries and spatial resolutions. Figure 13 shows these three datasets for the same Addis Ababa locale.

**Figure 13: Reference map polygons (left), the NYU 30 meter classification (center), and the GHSL38.2 meter classification (right**



In order to make comparisons of spatial agreement we opted to transform all datasets to the same spatial resolution so that we could make one-to-one comparisons across individual units. We chose to transform all datasets to the GHSL pixel resolution and pixel grid. For the *Atlas* classifications, this required resampling and reprojecting the data to the larger GHSL resolution and snapping the data to the GHSL pixel grid. For the reference map data, this required rasterizing and reprojecting the polygon data, snapping to the GHSL pixel grid, and assigning pixels a label corresponding to the majority class within the pixel space. Figure 14 shows the transformed datasets at the GHSL spatial resolution.

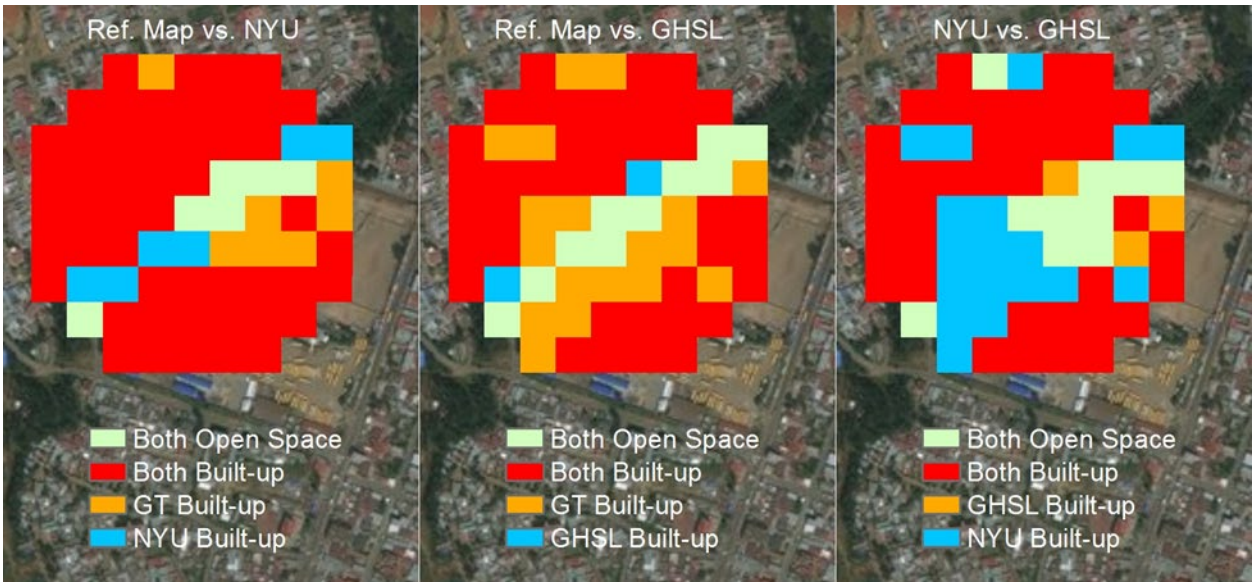
**Figure 14: The rasterized reference map data (left), the resampled NYU data (center), and the GHSL 38.2 meter classification (right).**

The information obtained from the comparisons was fed into an error matrix, or confusion matrix, which provides a formal representation of the agreement between classes and a basis for calculation of several measures of map accuracy including producer's accuracy and user's accuracy of the different classes, as well as overall map accuracy (Congalton and Green 2009).

Pixel Based Assessment

The first measure of accuracy assessed agreement at the pixel level. In each city, the pixel level comparisons for all pixels in all locales were pooled to populate a city specific error matrix. Figure 15 illustrates the four outcomes associated with pixel level comparisons. There can be agreement of open space, indicated by green, agreement of built-up, indicated by red, omission errors of the built-up class, indicated by orange; areas the reference map indicated to be built-up but that the *Atlas*/GHSL datasets identified as open space, and commission errors of the built-up class, indicated by blue, areas where the reference map indicated open space, but *Atlas*/GHSL datasets identified as built-up. The *Atlas* vs. GHSL comparison on the far right is not an assessment of accuracy per se, but an assessment of agreement.

**Figure 15: Comparisons at the pixel level for the reference data vs. NYU, reference map vs. GHSL, and NYU vs. GHSL.**




Locale Based Assessments

Whereas the pixel-based assessment describes the the spatial agreement of pixel labels across the different maps, the locale-based assessments, of which there are two, describe different measures of quantity agreement.

Following Potere et al (2009), the first locale-based assessment assigned one of two class labels, built-up or open space, to individual locales within a city. The locale label was determined by the majority class within the locale, obtained via the aggregation of pixel values. Comparisons at the locale level between the reference map locale label and the map classification locale label were used to populate error matrices to produce the standard accuracy measures.

The second locale-based assessment was designed to be less rigid than the all-or-nothing majority class label. In the all-or-nothing approach, locale comparisons of 45 percent and 55

percent built-up, only a 10 percent difference, are treated as disagreement while locale comparisons of 55 percent and 95 percent –a 40 percent difference, are treated as agreement. For each individual locale we compared the percentage built-up in reference data versus the map classification. The mean difference in percent built-up across all locales is a measure of the quantity agreement between the reference map data and the map classification. Whereas overestimates and underestimates across locales may cancel each other out, suggesting high overall agreement in percent built up, we can learn about the total error associated with each the maps by observing mean absolute difference in percent built-up. There is no confusion matrix associated with this type of measure.

**Map Comparisons**

While comparing the *Atlas* and GHSL classifications to an independent reference map to assess their accuracy is important for understanding how the two datasets differ, direct comparison between the *Atlas* and GHSL datasets is also important for understanding how and why map accuracy matters. The relationship between map accuracy and intermap agreement is unclear, particularly with respect to derived outputs from the classifications. In other words, it may be possible for the size and spatial agreement of urban extents created by the *Atlas* and GHSL datasets to be more similar – or different – than the pixel-based assessments of accuracy suggest. This is an empirical question we will address. We are interested in exploring intermap agreement at various spatial scales, from the pixel level all the way up to the urban extent.

Pixel and Locale Based Comparisons

Following the approach of the accuracy assessment, we explore pixel-based and locale-based agreement between the *Atlas* and the GHSL datasets. There is no reference map, or ground truth, so to speak, so the comparison does not provide a measure of accuracy, only of agreement. The pixel-level comparison is based on pooling pixels at two levels: all pixels in all cities for a global assessment and all pixels at the city level for a city-based assessment for all cities. Similarly, locales are pooled across all sample cities and within all cities. Locale based comparisons are based on the majority class locale label and the percent built-up within the locale.

Urban Extent Comparisons

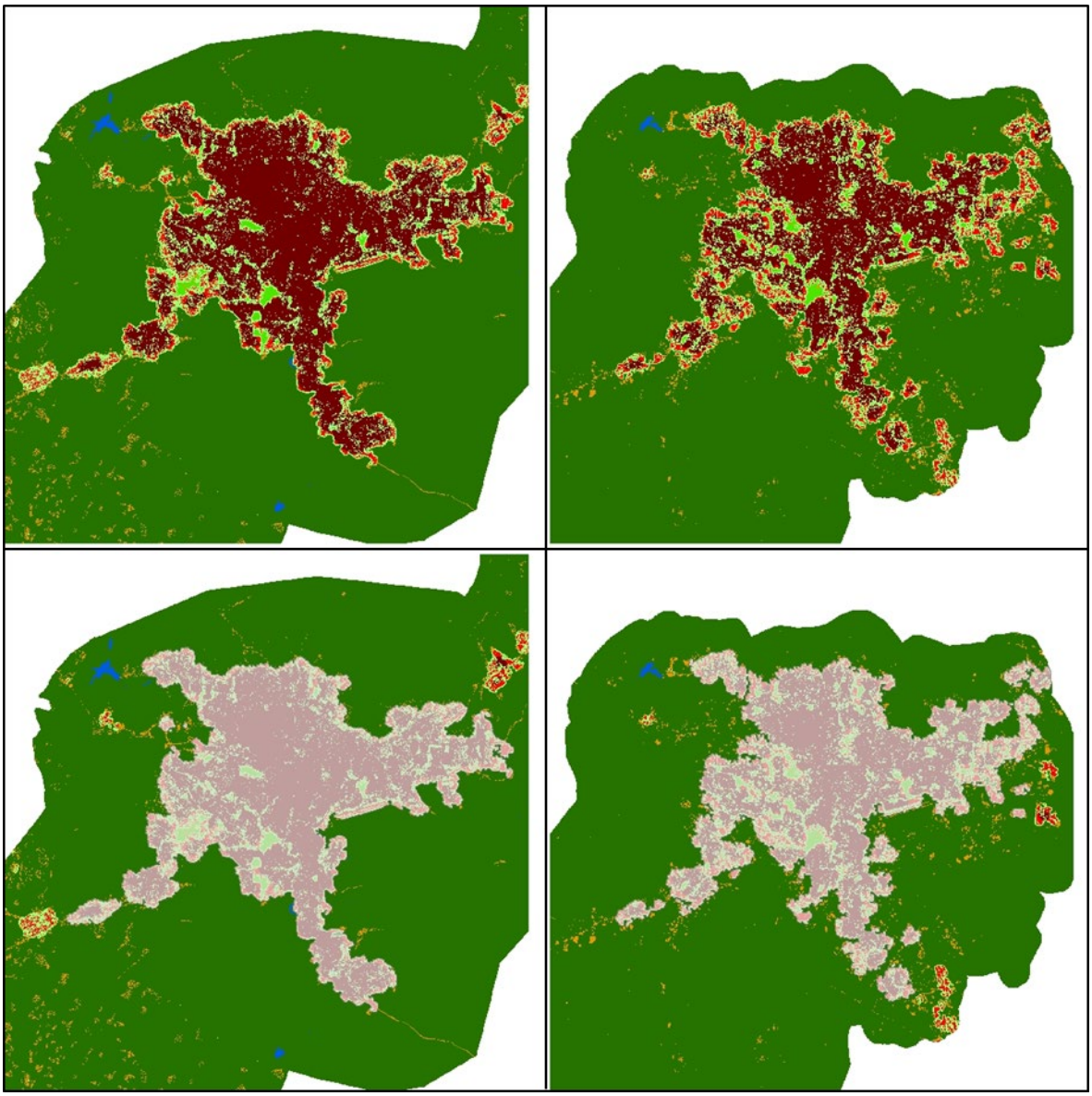
We are interested in comparing the derived outputs of the classifications, or the urban extents because of their clear connection to monitoring of city-level indicators.

To create urban extents from the GHSL data, we modified our computer scripts to run the landscape analysis and clustering rules on the 38.2 meter resolution GHSL data. We also resampled the *Atlas* classifications to the GHSL spatial resolution and recreated urban extents to facilitate spatial comparisons between the two datasets. A side-by-side comparison of the 7-way classification based on the *Atlas* and GHSL datasets for the area surrounding Addis Ababa, including the three subcategories of built-up, three subcategories of open space, and water, is presented in the top row of figure 16. The *Atlas* data is displayed on the left side and the GHSL data on the right side. The density of built-up area in the GHSL dataset appears somewhat sparser than the *Atlas* dataset, which appears more tightly packed. The bottom row shows the superimposed urban extent boundary generated by each dataset on top of the 7-way

classifications. Even though the *Atlas* data precedes the GHSL data by approximately three years, the size and shape of the urban extents show to have a high degree of correspondence.

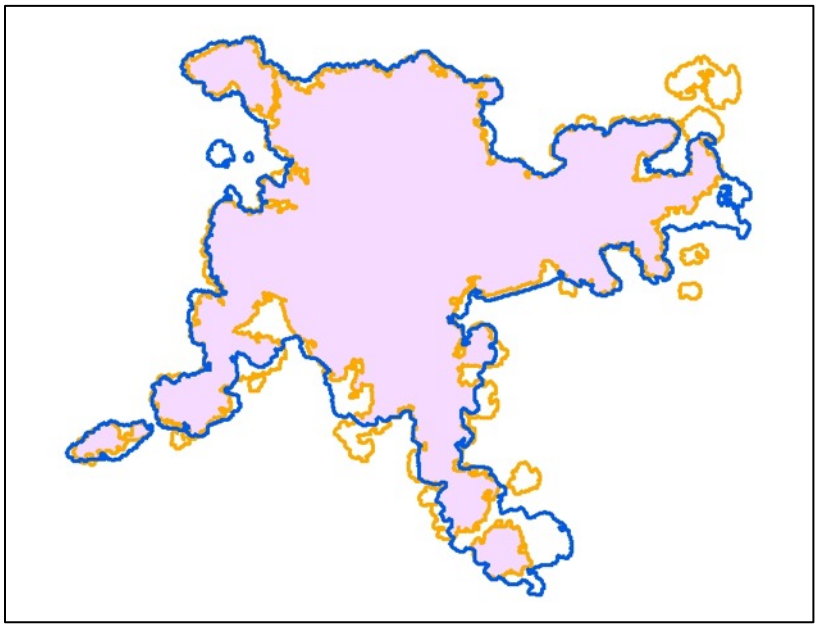
We can quantify the spatial agreement between the two urban extents by superimposing one urban extent on the other and calculating the area shared by both extents, and the area exclusive to the *Atlas* extent and the area exclusive to the GHSL extent. In figure 17, we see the outline of the *Atlas* extent in blue the outline of the GHSL extent in orange. The shared area is shaded in purple.

**Figure 16: The Landscape Analysis tool applied to the *Atlas* and GHSL classifications (top row) where *Atlas* data is shown on the left and GHSL data on the right. The urban extent boundary generated by each dataset (bottom row).**



The *Atlas* extent had an area of 292 km$^2$ compared to an area of 279 km$^2$ for the GHSL extent. From the perspective of the *Atlas* extent, the GHSL output was 4.5 percent smaller. The shared area between the two extents was 255 km$^2$. This means 87 percent of the *Atlas* area was shared with GHSL and 91 percent of the GHSL area was shared with the *Atlas* Approximately 37 km$^2$ of the *Atlas* extent was not shared area (13 percent) and 24 km2 of the GHS area was not shared area (9 percent).

**Figure 17: The Addis Ababa circa 2014 outlines of *Atlas* urban extent (blue), the GHSL urban extent (orange), and the shared area of the two urban extents (purple).**



While the accuracy comparisons rely on digitization of reference imagery, which could only be obtained for the most recent time period, it is possible to compare urban extents created by the *Atlas* and GHSL datasets at all three analysis periods: 1990, 2000, and 2014. We report on the quantity and spatial agreement of urban extents across all time periods.

## Results

### Accuracy Based on Pooled Data

Pixel-Based Measures

We first pooled all 1,769,740 pixels across all cities to populate *Atlas* and GHSL specific error matrices. The interpreted accuracy measures are displayed in Table 2.

**Table 2: Accuracy measures based on pooling of all pixels in the two datasets.**

| | Atlas | GHSL |
|---|---|---|
| Overall accuracy | 77% | 78% |
| Producer's accuracy, built-up | 86% | 82% |
| Producer's accuracy, open space | 55% | 69% |
| User's accuracy built-up | 81% | 86% |
| User's accuracy open space | 63% | 63% |

Overall accuracies are across the two datasets are very similar, 77 percent for the *Atlas* and 78 percent for the GHSL. The breakdown of producer's and user's accuracy for the different classes reveals differences in the performance of the two datasets that the single accuracy measure obscures. Producer's accuracy for the built-up class is a measure of omission error and reflects the degree to which a pixel in the reference map is correctly identified in the produced map (the *Atlas* or GHSL classification). In 86 percent of cases where a pixel was built-up in the reference map, it was also identified as built up in *Atlas* and in 14 percent of cases, the *Atlas* failed to identify reference pixels as built-up, resulting in omission errors. The *Atlas* omits built up pixels in the reference only slightly less than the GHSL. The largest difference between the two datasets concerns producer's accuracy for open space. When there is an open space pixel in the reference map, GHSL correctly identifies that pixel as open space 69 percent of the time, while the *Atlas* only identifies it correctly 55 percent of the time. The two datasets omitted 31 percent and 45 percent of reference map open space pixels respectively. User's accuracy is a measure of commission errors and indicates the probability that a pixel category in the produced map actually represents that pixel category on the ground. When the *Atlas* identifies a pixel as built-up, it is actually built-up 81 percent of the time and in 29 percent of the time, that built-up designation is a false alarm, or a commission error. Built-up user's accuracy for GHSL is slightly higher, 86 percent. User's accuracy for open space across the two datasets is the same, 63 percent.

Locale-Based Measures

We then pooled all 16,764 locales across all cities to populate Atlas and GHSL specific error matrices, where the reference locale and the comparison locale were assigned majority class labels. The interpreted accuracy measures are shown below in table 3.

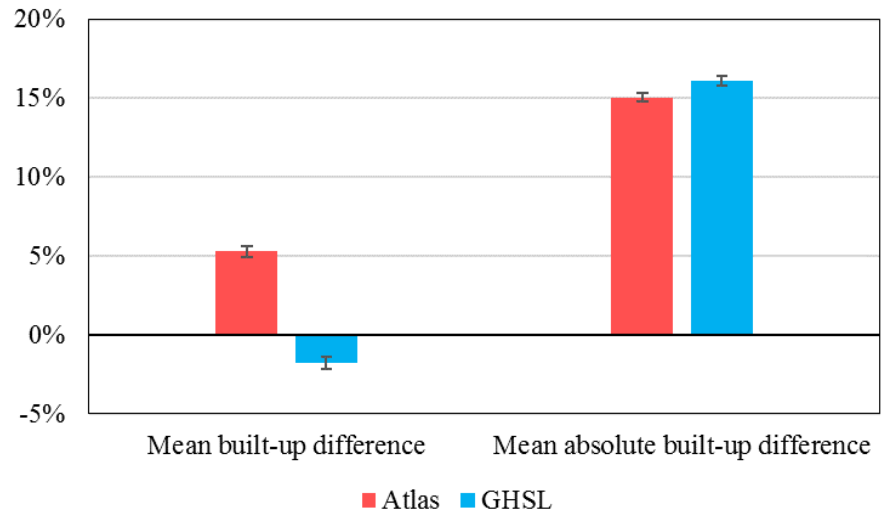
**Table 3: Accuracy measures based on pooling of all locales in the two datasets.**

| | Atlas | GHSL |
|---|---|---|
| Overall accuracy | 83% | 84% |
| Producer's accuracy, built-up | 93% | 86% |
| Producer's accuracy, open space | 56% | 77% |
| User's accuracy built-up | 85% | 91% |
| User's accuracy open space | 74% | 66% |

Compared to the pixel-based analysis, we see slightly higher values across all measures in both datasets. The overall accuracy of locales using *Atlas* data is 83 percent compared to 84 percent for GHSL. Again, producer's accuracy for the built-up class is higher for the *Atlas* while producer's accuracy for the open space class is higher for GHSL. The differences are somewhat larger compared to the pixel-based analysis. There is now a six percentage point difference between the two datasets for built-up producer's accuracy and a 19 percentage point difference for open space producer's accuracy. User's accuracy for the built-up class is still separated by 6 percentage points but user's accuracy for open space is 8 percentage points higher in the *Atlas* compared to GHSL, 74 percent vs. 66 percent. In the pixel-based assessment, user's accuracy for open space was 63 percent for both datasets. A possible explanation for the differences in user's and producer's accuracy for open space in the locale-based analysis is as follows: the *Atlas* may have identified fewer open space locales than GHSL, but the relatively fewer open space locales it identified were open space locales in the reference dataset at a higher rate than GHSL. The GHSL may have identified more open space locales, which allowed it correctly identify more open space in the reference data, but this may also have led an overidentification of open space, which resulted in a higher rate of false alarms, or commission errors.

The mean difference in percent built-up at the locale level is shown in figure 18. The mean difference for *Atlas* locales is positive, meaning they identify more built-up area on average than reference map locales. This mean difference is 5 percent and the median difference is 3 percent. The mean difference for GHSL is negative, meaning that on average the GHSL underestimates the percent built up compared to reference map locales. The mean GHSL difference is negative 2 percent and the median difference is zero. The mean difference pools overestimates and underestimates which can cancel each other out. A more accurate reflection of the error across locales is obtained by calculating the mean absolute difference in percent built-up. The mean absolute difference paints a different picture, showing a 15 percent difference for *Atlas* locales and a 16 percent difference for GHSL locales. One possible interpretation of this finding is that the overestimates and underestimates in the GHSL dataset cancel each other out more equally, bringing its mean difference closer to zero. The negative mean difference indicates that there are more underestimates than overestimates in the GHSL dataset. In the *Atlas,* it would appear that there is marginally less overall error but the errors do not cancel each other out as much as the GHSL dataset. Rather, there are more overestimates than underestimates, and this leads to a positive mean difference, further away from zero than the GHSL dataset.

**Figure 18: Mean difference in percent built up across all locales with 95 percent confidence intervals.**

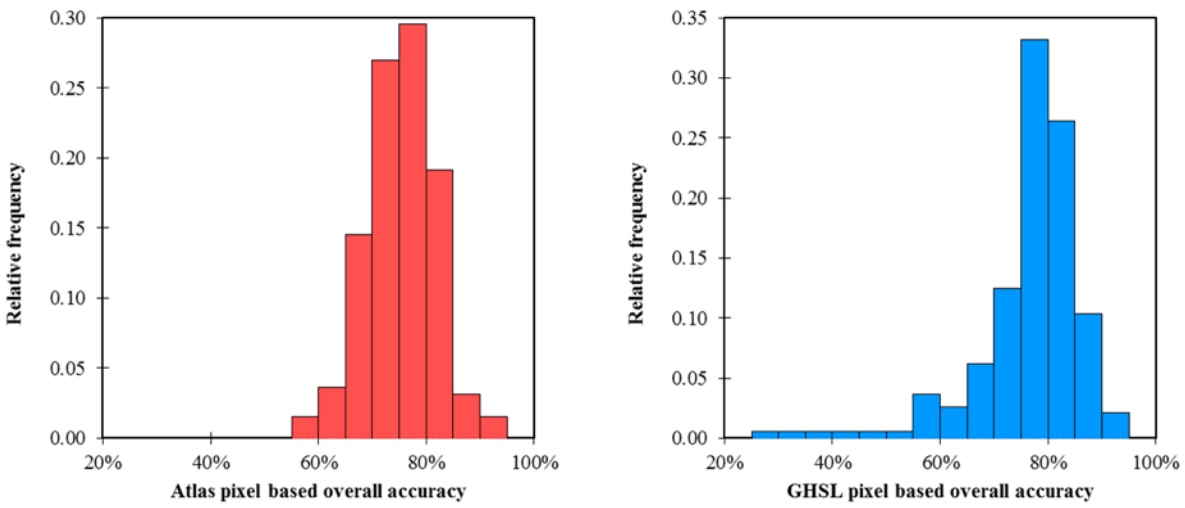



**Accuracy Based on City-Level Data**

Pixel-Based Measures

A second layer of the accuracy analysis aggregated pixels and locales at the city level to derive city specific accuracy measures. If the performance of the *Atlas* and GHSL maps were consistent across cities, we might expect all cities to have the same accuracy scores. Figure 19 shows the distribution of overall accuracy at the city level for the *Atlas* and GHSL datasets. While both distributions show a strong central tendency, the distribution of scores range between 57 percent and 92 percent for the *Atlas* and between 30 percent and 92 percent for the GHSL. Differences observed at the city level may be due to a variety of factors, including: poor performance of the classification procedures in particular climates or geographic settings, temporal differences between the reference map data and the comparison maps at the city level, reference map errors that are not randomly distributed but associated with specific cities, and the case of the *Atlas,* variation in the skills of analysts, that may have affected the quality of the classifications or the quality of the digitization and labeling of reference imagery.

**Figure 19: The distribution of pixel based overall accuracy calculated at the city level in the *Atlas* (left) and GHSL (right) datasets.**




Averaging city-level overall accuracy, we find that the two datasets perform very similarly, 75 percent for the *Atlas* and 76 percent for GHSL. A paired t-test reveals that there is no statistical difference in the average overall accuracy of the two datasets at the 95 percent confidence interval. Average overall accuracy is two percentage points lower for both datasets compared to the result obtained from pooling all pixels across all cities. Since the average city-based score is averaged across cities and since the number of pixels associated with each city is not constant, perhaps low scoring cities with relatively fewer pixels bring down the average compared to pooling all pixels for all cities together. The breakdown of producer and user accuracy, in table 4, shows a pattern very similar to pooled results.

**Table 4: Pixel based accuracy calculated by averaging across city level accuracy.**

| | Atlas | GHSL |
|---|---|---|
| Overall accuracy | 75% | 76% |
| Producer's accuracy, built-up | 85% | 78% |
| Producer's accuracy, open space | 54% | 69% |
| User's accuracy built-up | 79% | 85% |
| User's accuracy open space | 64% | 65% |

The *Atlas* has higher producer's accuracy for the built-up class but also lower user's accuracy for the built-up class. This is likely explained by overidentification of the built-up class. Overidentification would ensure that built-up in the reference map is correctly identified but it might also mean that the rate of false positives, or commission errors, associated with the built-up class may be high as well. Producer's accuracy for open space is 15 percentage points higher for the GHSL while user's accuracy for one space is essentially the same. When there is open space on the reference map, GHSL is less likely to omit that open space than the *Atlas*. When we look at individual open space pixels across the GHSL and *Atlas* classifications, those pixels

actually represent open space on the ground nearly two-thirds of the time.

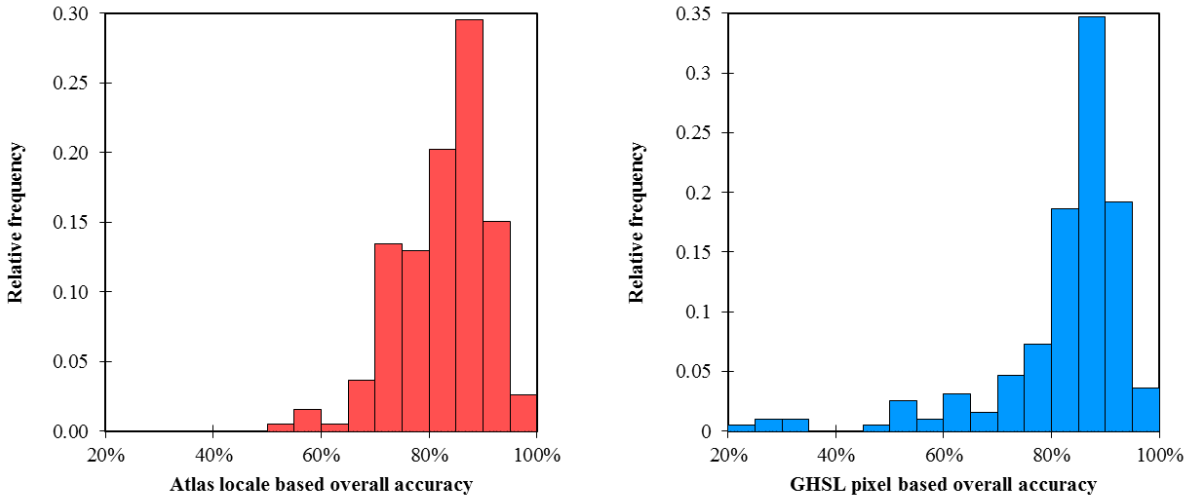
Locale-Based Measures

When we average locale-based accuracy measures across cities, we find minor differences in the results compared to those calculated by pooling all locales across all cities. This average masks variation in average locale based accuracy at the city level, shown in figure 20. While several cities have accuracies of 95 percent or higher there are also a number of poor performing cities. The breakdown of overall accuracy, and producer's and user's accuracy is shown in table 5.

**Table 5: Locale based accuracy calculated by averaging across city level accuracy.**

| | Atlas | GHSL |
|---|---|---|
| Overall accuracy | 82% | 82% |
| Producer's accuracy, built-up | 92% | 82% |
| Producer's accuracy, open space | 59% | 78% |
| User's accuracy built-up | 84% | 92% |
| User's accuracy open space | 74% | 68% |

**Figure 20: The distribution of locale based overall accuracy calculated at the city level in the *Atlas* (left) and GHSL (right) datasets.**

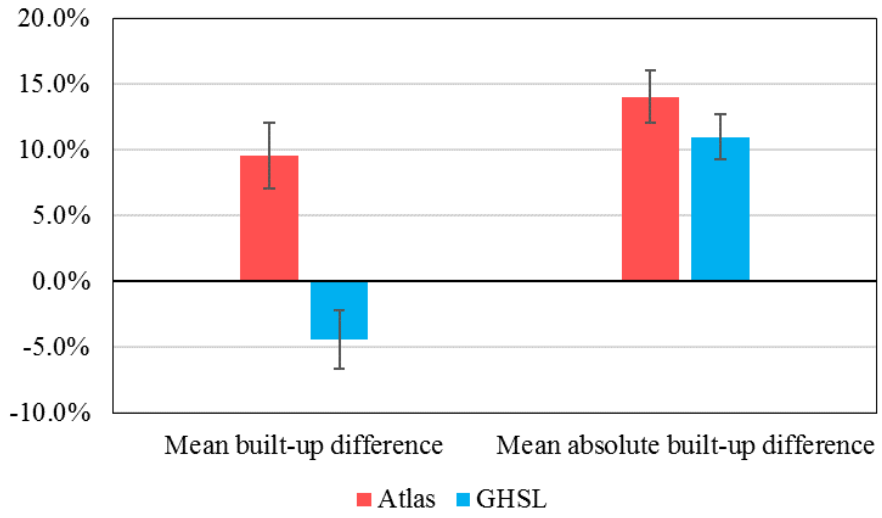



The overall accuracy of the two datasets is the same, 82 percent. Overall accuracy at the locale level is higher than overall accuracy at the pixel level and only one percentage point lower compared to pooling locales across all cities. Producer's accuracy for the built-up class is higher in the *Atlas* compared to GHSL, 92 percent to 82 percent, but user's accuracy for the built-up class is lower in the *Atlas* compared to GHSL, 84 percent to 92 percent. Again, this may be related to overidentification of the built-up class, which leads to correct identification of built up in the reference map but also a slightly higher rate of commission errors. Producer's accuracy

for open space is lower in the *Atlas* than the GHSL, 59 percent compared to 78 percent though performance for the two datasets is reversed for open space user's accuracy, 74 percent versus 68 percent. The *Atlas* identifies open space in the reference map less often than the GHSL but when the *Atlas* identifies open space, it is slightly more likely to actually be open space the ground.

The mean difference in percent built-up at the locale level, averaged across cities, is shown in figure 21. The mean difference for *Atlas* locales is positive, meaning *Atlas* locales identify more built up area on average than reference map locales; the negative mean difference for GHSL indicates that GHSL locales identify less built up area on average than reference map locales. The mean values are greater compared to the result obtained from pooling all locales across all cities. The *Atlas* mean difference is now 10 percent compared to 5 percent, and for the GHSL it is now negative 5 percent compared to negative 3 percent. These differences are likely explained by poorly performing cities that pull the pooled mean difference further away in a positive or negative direction.

**Figure 21: Mean difference in percent built up across locales across cities with 95 percent confidence intervals.**




The mean absolute built-up difference is 14 percent and 11 percent for the *Atlas* and GHSL datasets respectively. The overlapping confidence intervals for these averaged values suggests that the total percent error in each of the datasets is not significantly different at the 95 percent confidence level. Nevertheless, it appears that the overestimates and underestimates of percent built-up in the GHSL dataset cancel each out more and that it has slightly more underestimates. This is why its mean built-up difference is negative and closer to zero than the *Atlas.* The *Atlas* would appear to overestimate the percent built up more consistently, leading to a positive mean built-up difference that is further away from zero than the GHSL.

**Map Comparisons, *Atlas* vs. GHSL**

Pixel-Based and Locale-Based Comparisons

We now focus on the comparison of the *Atlas* classifications to the GHSL classifications. The

accuracy measures, which compared both datasets to reference map data, showed the same general trends whether the results were pooled or whether they were aggregated by city and averaged across cities. Results obtained by averaging accuracies across cities were typically lower by a slight margin. For the comparison of the *Atlas* to the GHSL we focus only on the pooled comparisons. Measures of agreement are shown below in Table 6.

**Table 6: *Atlas* vs. GHSL pixel-based and locale-based measures of agreement.**

| | Pixel-based | Locale-based |
|---|---|---|
| Overall agreement | 78% | 81% |
| Producer's agreement, built-up | 88% | 94% |
| Producer's agreement, open space | 56% | 52% |
| User's agreement built-up | 80% | 81% |
| User's agreement open space | 71% | 80% |

Although this comparison is not an assessment of accuracy, we retain the concepts of producer's accuracy and user's accuracy, replacing the word accuracy with agreement. The *Atlas* classifications are the comparison map while the GHSL is the reference map. The overall pixel-based agreement shows 78 percent of all pixels were assigned the same class label across the two datasets. Locale based agreement, based on a majority classifier, was slightly higher, 81 percent. Built-up pixels in the GHSL dataset were correctly identified by the *Atlas* 88 percent of the time and built-up locales were correctly identified at an even higher rate of 94 percent. There is much lower agreement with respect to open space. When the GHSL identifies an open space pixel, the *Atlas* identifies that pixel as such only 56 percent of the time, and the outcome is even poorer, 52 percent, at the locale level. User's agreement for built up, or how often a built-up pixel or locale in the *Atlas* corresponds to built-up in the GHSL dataset is consistent across pixels and locales, 80 and 81 percent respectively. User's agreement for open space was lower for pixels but approximately the same for locales. The results suggest that there is considerable disagreement in the identification of open space class between the two datasets but much better agreement for the built-up class. The *Atlas* failed to identify between 48 and 44 percent of open space identified by GHSL, depending on whether the measure is pixel or local based.

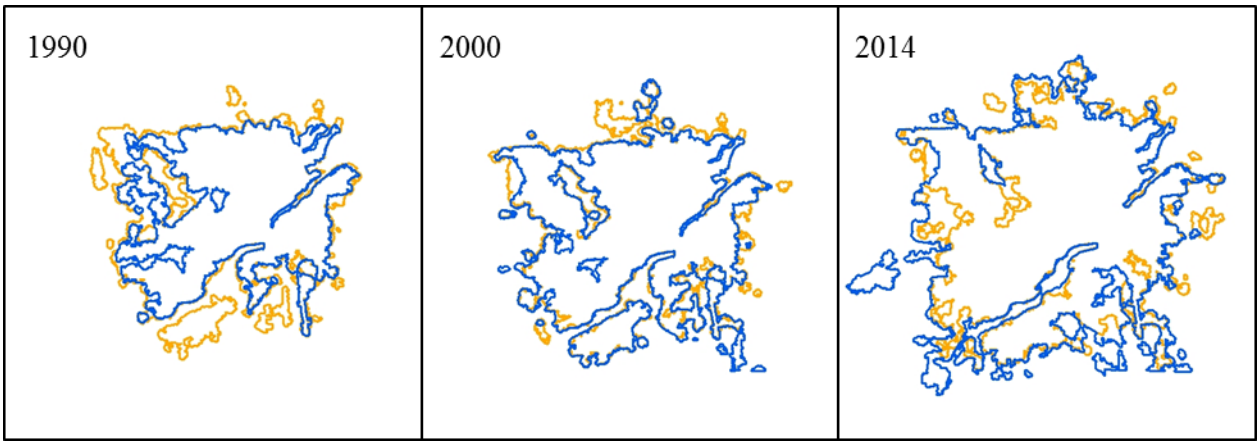
The mean difference in percent built-up between the *Atlas* and GHSL across all locales is seven percent and the median difference is zero. The mean difference is pulled upward by some *Atlas* locales that identify much more open space than GHSL. The mean absolute difference in percent built-up between locales is 17 percent. In other words, sometimes the *Atlas* identified more built up area than GHSL and sometimes it identified less, but overall, the positive differences appear to outweigh the negatives, evidence by the mean difference of positive seven percent.

Urban Extent

The Toledo, Ohio urban extents created with the GHSL dataset and with the resampled *Atlas* data (at the GHSL spatial resolution) are shown in figure 22. Visual inspection suggests relatively good size and spatial agreement across the three time periods. In this particular example we find that the GHSL extent was 36 percent larger than the *Atlas* extent in 1990, 9 percent larger in 2000 and 11 percent smaller in 2014. The Toledo example is representative of

the general trend we observe across all cities, namely, a larger GHSL extent in 1990, a more similarly sized extent in 2000, and a larger *Atlas* extent in 2014.

**Figure 22: The outlines of the Toledo, Ohio urban extent created by the *Atlas* (blue) and the GHSL (orange).**




The average percent difference in the size of the urban extent, averaged across 182 cities, is shown in table 7. Unresolved data processing errors resulted in the exclusion of 12 cities that were included in the accuracy analysis. The percent difference calculation shows the size of the GHSL extent relative to the *Atlas* extent.

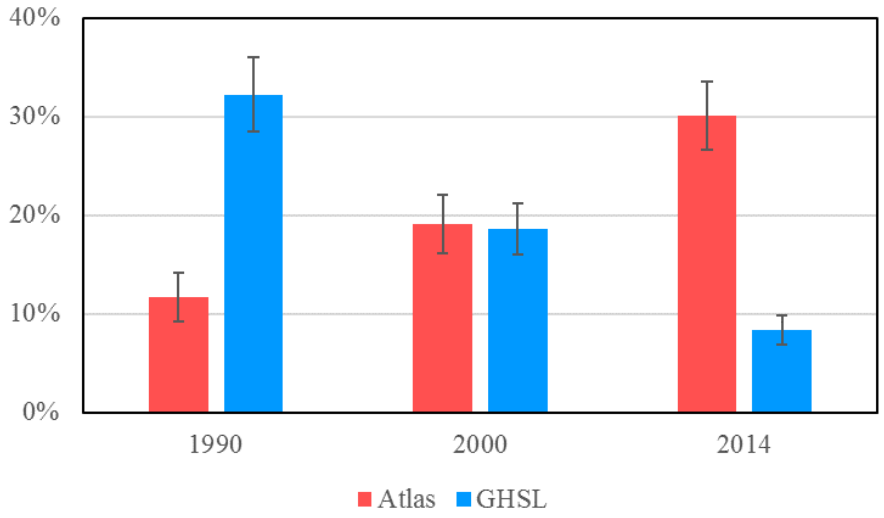
**Table 7: Percent difference in urban extent created by the *Atlas* and GHSL datasets.**

| | 1990 | 2000 | 2014 |
|---|---|---|---|
| Average | 100% | 14% | -21% |
| Lower 95% CI | 59% | 3% | -26% |
| Upper 95% CI | 141% | 26% | -16% |
| Median | 25% | 0% | -17% |

The average result for 1990, 100 percent, is striking. Extents created with GHSL data are on average twice as large as extends created with *Atlas* data. This large value can be explained by the presence of extreme outliers in 1990. The two most extreme outliers, Kozhikhode, India and Rawang, Malaysia, have extents more than 1,000 percent larger than their counterpart *Atlas* extents. The median 1990 GHSL extent is only 25 percent larger. In 2000, the difference was significantly smaller compared to 1990, only a 14 percent average difference between GHSL extents and *Atlas* extents. Although the 95 percent confidence interval for the year 2000 average is above zero, the median percent difference value across cities is zero, indicating that just as many GHSL extents are larger than *Atlas* extents as they are smaller. The positive average value suggests that when GHSL extents are larger than *Atlas* extents that difference is much greater than when they are smaller. In 2014, the relationship between GHSL and *Atlas* extents reverses. GHSL extents were found to be on average 21 percent smaller than *Atlas* extents and that difference is significant.

We explore the spatial agreement between urban extents created by each dataset by looking at the share of each urban extent that is exclusive to that dataset, meaning area that is not shared by the urban extent created by the other dataset. This area corresponds to the non-shaded areas in the Addis Ababa image in figure 17. The average shared area of a dataset is 1 minus its exclusive area. The average shares of exclusive areas are shown in figure 23.

**Figure 23: The average share of area exclusive to the urban extent generated at each time period with 95 percent confidence intervals.**




In 1990, 32 percent of the area of GHSL extents was exclusively GHSL area and 68 percent of GHSL areas were shared with *Atlas* extents. The exclusive area is smaller for *Atlas* extents in 1990, only 12 percent, which is expected since *Atlas* extents were smaller than GHSL extents at that period, increasing the chance that they lie within GHSL extents. In 2000 the average share of exclusive urban extent area for each dataset was 19 percent. The size agreement of extents across datasets (14 percent) would appear to be more similar than the spatial agreement of extents at this period. In 2014, we observe a reversed situation compared to 1990 due to the smaller size of GHSL extents relative to *Atlas* extents. Approximately 30 percent of the area of *Atlas* extents is not shared with the GHSL extents while only 8 percent of GHSL areas are not shared by *Atlas* extents.

## Discussion

### Accuracy

It is encouraging that the overall accuracies obtained for the *Atlas* and GHSL datasets are very similar for the 2014 period. Pooled results show a difference in overall accuracy of only one percentage point. When the pixel and locale data is aggregated at the city level, we find that the average paired difference in overall accuracy across the two datasets is not significantly different than zero. A singular focus on overall accuracy masks differences in the accuracies of the built-up and open space classes, however. For the 2014 period we observed that the *Atlas* identified the built-up class in the reference data somewhat better than the GHSL but that this was also

associated with a general pattern of over-identification that led to a higher rate of false alarms for the built-up class. The GHSL dataset identified the open space class in the reference data better than the *Atlas*, but when each dataset claimed a locale to be open space, the *Atlas* performed slightly better, perhaps because it was more parsimonious in its open class assignment. These differences are important for understanding why two datasets with virtually identical overall accuracies are associated with urban extents whose sizes are significantly different from each other.

It is worth noting that the accuracy measures obtained from this analysis are lower than other reported accuracies of Landsat classifications in globally distributed urban sites. Angel et al (2005) used a virtually identical classification procedure and obtained an average overall accuracy of 89.2 percent; Schneider and Woodcock (2008) obtained accuracies between 84 and 97 percent; Potere et al (2009) obtained an average overall accuracy of 87.1 percent; and Pesaresi et al (2016) obtained accuracies on the order of 90 – 97 percent. While Leyk et al (2018) assessed the GHSL accuracy at different time periods in the United States and a report variety of accuracy measures, they do not report a comparable overall accuracy measure.

A direct comparison of the accuracy figures may be somewhat misleading owing to differences in the way accuracy was assessed. Angel et al (2005) and Schendier and Woodcock (2008) randomly sampled one-pixel sites across city study areas and analysts visually interpreted high-resolution satellite imagery to assign the pixels a binary reference class label. In Potere et al (2009), analysts assigned majority class labels to 0.132 km2 hexagonal areas, similar in size to our locales, based on the photo-interpretation of Google Earth imagery (Potere, 2008). Pesaresi et al (2016) did not use photointerpretation of pixels but compared the GHSL built-up grid against reference data primarily from Europe and the United States.

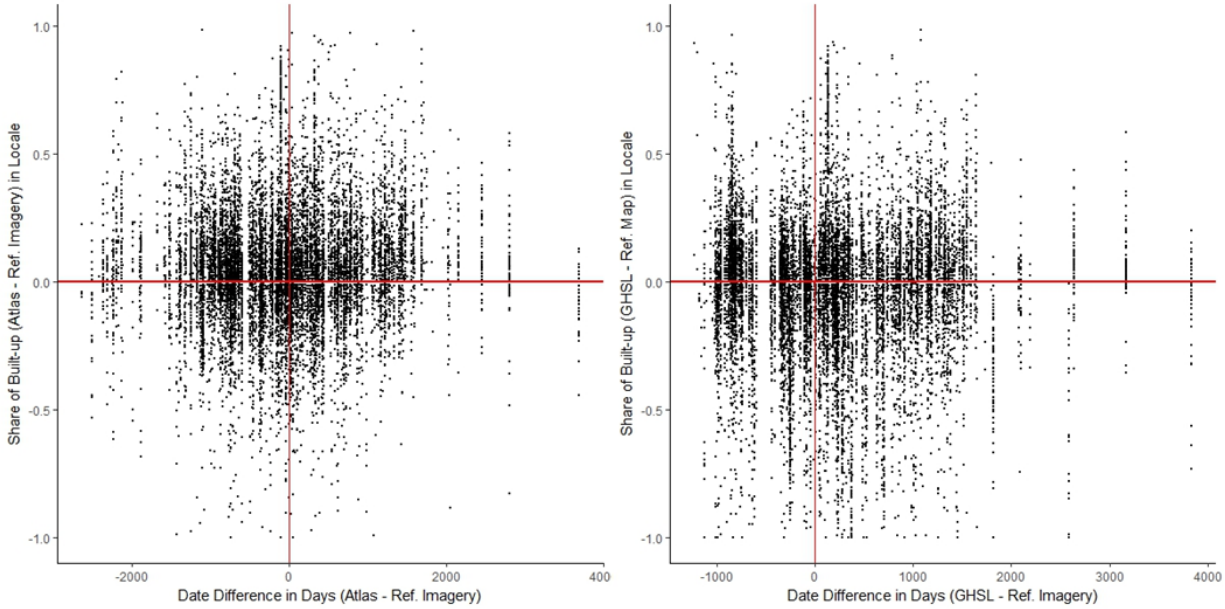
Our accuracy assessment relied on a globally distributed reference dataset that was created by the manual digitization of high resolution satellite imagery and the labeling of digitized polygons by analysts. In other words, our reference dataset was different in many respects. Furthermore, this dataset was rasterized to allow for one-to-one comparisons with the GHSL dataset. Perhaps an accuracy assessment based on the random sampling of one-pixel sites across study areas, where reference pixel labels are determined by the photointerpretation of individual pixels into binary ‘majority built’ or ‘majority open space’ classes might lead to a result similar to those observed in other studies. We may conduct exploratory assessments in a small sample of cities to understand how the results associated with the two methods compare.

At least four factors may have affected the accuracies we observed: (1) temporal relationships between the reference map imagery and the Landsat imagery, (2) reference map digitization or labeling errors, (3) information loss when the reference map polygon data is rasterized and resampled to match the GHSL resolution and when the *Atlas* data is resampled to match the GHSL resolution, and (4) idiosyncracies unique to the *Atlas* and GHSL classification methods.

There would appear to be very little effect of temporal relationships on explaining error. In figure 24, the temporal relationship between the comparison map date and the reference imagery date is represented by the x-axis and the share of built-up in a comparison map locale minus the share of built up in a reference map locale is represented by the y-axis. Each dot represents a locale.

Observations to the right of zero on the x-axis represent locales where the comparison map date comes after the reference imagery date. Observations to the left of zero on the x-axis represent locales where the comparison map date precedes the imagery date. If the reference imagery date precedes the comparison map date (observations to the right of zero on the x-axis), we might expect the reference data to say there is less built-up area while the comparison map says there is more built-up area. This would lead to more observations in the upper right quadrant. Conversely, if the reference image date comes after the comparison map, we might expect the reference data to say there is less built-up area while the comparison map says there is more built up area. This would lead to more observations in the lower left quadrant. If most observations were in the upper right and lower left quadrants, then the temporal relationships between reference imagery and the comparison maps would help explain differences in the amount of built up in locales. Instead, we see a random distribution of points across the four quadrants in both datasets, suggesting that differences between the amount of built in comparison map locales and reference map locales is not related to timing.

**Figure 24: Temporal relationships in the datasets on the x-axis vs. differences in the amount built up in locales on the y-axis.**



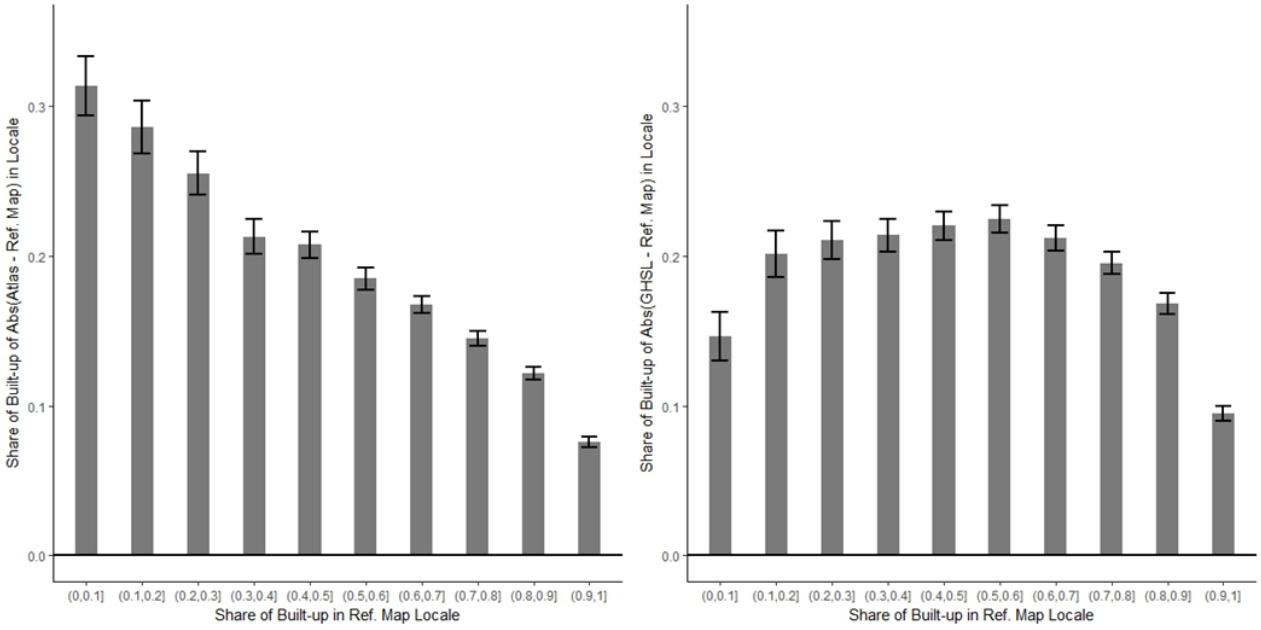

We were unable to explore potential errors in reference map digitization and labeling in a systematic way, but the inspection of a number of misclassified pixels and locales suggests that it is an issue that needs to be taken seriously in interpreting the results. Since analysts were tasked with digitizing and labeling blocks, mixed block spaces pose a problem since they may only receive a single label. Figure 25 illustrates an example of this problem in a locale in Halle, Germany. The image on the left shows all block outlines that were assigned one of the labels discussed in section 2.2.3. The image on the right highlights those blocks identified as open space by the analyst. While the open space bocks are clearly open space, there also appear to be some open space patches distributed across other blocks that have a mixed built-up/open space character. This locale exemplifies the pattern of overall errors observed in the *Atlas* dataset, namely lower producer's accuracy in identifying open spaces in the reference map, but lower

commission errors associated with *Atlas* open space areas.

**Figure 25: The digitalization of locale polygons in Halle, Germany, and the problem of mixed block spaces.**

We were unable to quantify the amount of information loss that occurs when reference map polygons are rasterized or when *Atlas* landcover classifications are resampled to the GHSL resolution. This procedure clearly leads to some degree of spatial generalization. This is visible from a comparison of figures 13 and 14. We do have an empirical insight into idiosyncrasies of the *Atlas* and GHSL classification methods and their proclivities for detecting built up area. Figure 26 shows the share of built up area in reference locales on the x-axis and the difference in the share of built-up in comparison locales and reference locales on the y-axis. Two different patterns are visible the two datasets.

**Figure 26: The share built up in reference locales on the x-axis vs. the difference in the share built in comparison locales and reference locales on the y-axis (*Atlas* on the left, GHSL on the right).**




Intuitively, it would seem easier for detection methods to correctly identify locales that were either mostly built-up or mostly empty. This what we see on the GHSL chart on the right, where reference locales with between 90 percent and 100 percent built-up are associated with the lowest difference in the share built-up on the y-axis, and reference locales with between 0 and 10 percent built up are associated with the second lowest difference in the share built-up on the y-axis. The increasing then decreasing hump like pattern in the GHSL is absent from *Atlas* locales. While reference locales that have a high share built up are associated with small differences, the difference increases linearly as the amount of built up in the reference locale decreases.

## Urban Extent Comparisons

It is difficult to reconcile differences in the sizes of urban extents created by the GHSL and the *Atlas* with the knowledge that the overall pixel-based and locale-based accuracies of the two datasets are essentially equal. It is also difficult to understand why we observe larger *Atlas* extents in 2014 but larger GHSL extents in 1990 and 2000. If the classification method in each dataset is consistent, we should expect to see the same relationship between extents over time.

With regard to the most recent time period, we have observed that the *Atlas* tends to identify more built-up area than the GHSL. This is evidenced by the higher producer's accuracy and lower user's accuracy for the built-up class, the positive mean difference for the share of built-up in locales seen in figures 18 and 21, and the trend observed in figure 26. A dataset that identifies more built-up pixels will also identify larger clusters and larger extents, all things being equal.

Even though the *Atlas* identified more built-up area in 2014 on average than the GHSL, we find that the difference in urban extents is typically greater than the difference in amount of built up area. In other words, the difference in urban extent is an accentuated version of the difference in built-up area. We explored this relationship by examining the level of saturation in an area defined by the union of urban extents created by each dataset in a given city. Saturation is simply the built-up area divided by total area. In this case, the total area is the union of the two urban extents. Across this unioned area, we have the set of *Atlas* pixels and we have the set GHSL pixels, and we can calculate the saturation associated with the two datasets, which is a number between 0 and 1. For each city, we calculated a saturation ratio, or the GHSL saturation divided by *Atlas* saturation. For each city we have also calculated the ratio of the GHSL extent to the *Atlas* extent. These ratios are shown in table 8.

**Table 8: Saturation ratios vs. urban extent ratios.**

| | 1990 | 2000 | 2014 |
|---|---|---|---|
| Average Saturation Ratio (GHSL ÷ Atlas) | 1.55 | 1.10 | 0.91 |
| Average Urban Extent Ratio (GHSL ÷ Atlas) | 2.00 | 1.14 | 0.79 |

Across all three time periods, the dataset that identified more built-up pixels in the unioned area, or the more saturated dataset, was associated with larger urban extents. In 1990, GHSL was 55 percent more saturated than the *Atlas* but GHSL extents were on average 100 percent larger. In 2014, GHSL was 91 percent as saturated as the *Atlas* across the unioned area but GHSL extents were only 79 percent as large. In 2000, GHSL was 10 percent more saturated than the *Atlas,* but GHSL extents were 14 percent larger.

Conceptually, urban extent should be more than a linear function of saturation; the spatial distribution of built-up pixels across a given area will also influence how built-up area and open space combine to create urban extents. Could changing spatial forms over time be contributing to the changing relationship between saturation and urban extent over time? Perhaps, but we are unable to answer this question definitively at this time. Further explorations of the data are needed to provide insights.

While we can quantify the accuracy of the *Atlas* and GHSL classifications at 2014, we are less certain about the datasets' accuracy at earlier periods. If the classification methods and input data quality are relatively consistent across time, we would expect the accuracy at 2014 to carry over to the earlier time periods. The large differences in saturation and urban extent at the 1990 are rather surprising and calls into question the accuracy of the datasets at that period. Additional scrutiny of the 1990 classifications may be necessary to resolve questions about historical trends.

## Conclusion

Comprehensive accuracy analyses of global built-up area datasets are very costly and we were fortunate that the manually digitized satellite imagery from second volume of the *Atlas* could be employed for this task. We have obtained an answer to our initial question about the accuracy of

the landcover classifications in the *Atlas* and GHSL datasets. When compared to an independent reference dataset, the *Atlas* and GHSL were found to have nearly identical overall accuracy but different accuracies for the built-up and open space classes. We also know how urban extents created by the two datasets compare. The extents created by the two datasets were significantly different in size within a given time period and the relationship between them varied across time periods: *Atlas* extents were larger in 2014 but GHSL extents were larger in 2000 and 1990.

While the knowledge gained is important to our understanding of the strengths and weaknesses of the GHSL and the *Atlas*, it has not provided a conclusive verdict on either dataset. It is difficult, but not impossible, to reconcile the nearly identical and relatively high accuracy in both datasets with urban extents that are significantly different in size. Perhaps these differences are the reflection of an urban extent methodology that is too sensitive to variations in the quantity and spatial distribution of built-up and open space pixels. It may be prudent to revisit and revise and urban extent methodology to make it more robust.

It may also be worthwhile to conduct a pixel-based accuracy assessment using high resolution satellite imagery in a sub-sample of cities. The Bing Aerial Imagery Analyzer for OpenStreetMap could be used to evaluate temporal relationships between the comparison map data and the reference imagery. We could then compare the accuracy obtained from the pixel-based assessment against the accuracy obtained from this analysis. We suspect the pixel-based accuracies will be higher owing to the problem of mixed blocks discussed in section 5. The extent to which mixed blocks or other digitization errors affect accuracy can be determined empirically. In a related vein, the feasibility of implementing an open sourced worldwide (quality controlled) assessment framework, similar to the one employed by Potere (2008), is worth exploring.

## References


Angel, Shlomo, Stephen Sheppard, and Daniel L. Civco. 2005. *The dynamics of global urban expansion.* Washington D.C.: Transport and Urban Development Department, World Bank.

Angel, Shlomo, Alejandro M. Blei, Jason Parent, Patrick Lamson-Hall, Nicolas Galarza, Daniel L. Civco, Qian Lei, and Kevin Thom. 2016. *Atlas of Urban Expansion—2016 Edition*, Volume 1: Areas and Densities, New York: New York University, Nairobi: UN-Habitat, and Cambridge, MA: Lincoln Institute of Land Policy

.

Angel, Shlomo, Patrick Lamson-Hall, Manuel Madrid, Alejandro M. Blei, Jason Parent, Nicolas Galarza, Kevin Thom. *Atlas of Urban Expansion—2016 Edition*, Volume 2: Blocks and Roads, New York: New York University, Nairobi: UN-Habitat, and Cambridge, MA: Lincoln Institute of Land Policy.

Blei, Alejandro, Shlomo Angel, Daniel L. Civco, Nicolas Galarza, Achilles Kallergis, Patrick Lamson-Hall, Yang Liu, and Jason Parent. 2018. Urban Expansion in a Global Sample of Cities: 1990 – 2014. Working Paper. Cambridge, MA: Lincoln Institute of Land Policy.

Congalton, Russell 1991. "A Review of Assessing the Accuracy of Classificaitons of Remotely Sensed Data." *Remote Sensing of the Environment*, *37*: 35-46.

Congalton, Russell and Kass Green. 2009. *Assessing the Accuracy of Remotely Sensed Data: Principles and Practices – Second Edition.* Boca Raton, FL: CRC Press.

Corbane, Christina, Martino Pesaresi, Panagiotis Politis, Vasileios Syrris, Aneta J. Florczyk, Pierre Soille, Luca Maffenini, Armin Burger, Veselin Vasilev, Dario Rodriguez, Filip Sabo, Lewis Dijkstra, and Thomas Kemper. 2016. "Big eart data analytics on Sentinel-1 and Landsat imagery in support to global human settlements mapping." *Big Earth Data, 1*(1-2), 118 – 144.

Galarza Sánchez, Nicolas, Yang, Liu, Shlomo Angel, and Kevin Thom. 2018. The 2010 Universe of Cities: A New Perspective on Global Urbanization. Working Paper, Cambridge MA: Lincoln Institute of Land Policy.

Pontius, Robert Gilmore and Marco Millones. 2011. "Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment." *International Journal of Remote Sensing, 32*(15): 4407 – 4429.

Klotz, Martin, Thomas Kemper, Christian Geiss, Thomas ESch, and Hannes Taubenböck. 2016. "How Good is the map? A multi-scale cross-comparison framework for global settlement layers: Evidence from central Europe." *Remote Sensing of Environment, 178:* 191 – 212.

Leyk, Stefan, Johannes H. Uhl, Deborah Balk, and Bryan Jones. 2018. "Assessing the accuracy of multi-temporal built-up land layers across rural-urban trajectories in the United States." *Remote Sensing of Environment, 204:* 898 – 917.

Pesaresi, Martino, Christina Corbane, Andreea Julea, Aneta J. Florczyk, Vasileios Syrris and Pierre Soille. 2016. "Assessment of the added-value of Sentinel 2 for detecting built-up areas." *Remote Sensing, 8*(4), 299.

Pesaresi, Martino, Michele Melchiorri, Alice Siragusa, and Thomas Kemper. 2016a. Atlas of the Human Planet 2016: Mapping Human Presence on Earth with the Global Human

Settlement Layer. JRC Science for Policy Report. European Commission, Joint Research Center: Luxembourg.

Pesaresi, Martino, Daniele Ehrlich, Stefano Ferri, Aneta J. Florczyk, Sergio Freire, Matina Halkia, Andreea Julea, Thomas Kemper, Pierre Soille, and Vasileios Syrris. 2016c. Operating procedure for the production of the global human settlement layer from Landsat data of the epochs 1975, 1990, 2000, and 2014. JRC Technical Report, European Commission.

Potere, David and Annemarie Schneider. 2007. "A Critical Look at Representations of Urban Areas in Global map." *GeoJournal*, *69*(1-2): 55 – 80.

Potere, David. 2008. "Horizontal positional accuracy of Google Earth's high-resolution imagery Archive." *Sensors*, *8*: 7973–7981.

Potere, David, Annemarie Schneider, Shlomo Angel, and Daniel Civco. 2009. "Mapping Urban Areas on a Global Scale: Which of the Eight Maps Now Available is More Accurate?" *International Journal of Remote Sensing, 30*(24): 6531 – 6558.

Roy, D.P., Wulder, M.A., Loveland, T.R., Woodcock, C.E., Allen, R.G., Anderson, M.C., Helder, J.R., Irons, D..M. **and others.** 2014. "Landsat-8: Science and Product for Terrestrial and Global Change Research." *Remote Sensing of Environment*, *145*: 154 – 172.

Schneider Annemarie, and Curtis Woodcock. 2008. "Compact, Dispersed, Fragmented, Extensive? A Comparison of Urban Expansion in Twenty-Five Global Cities Using Remotely Sensed Data, Pattern Metrics, And Census Information." *Urban Studies, 45:* 659-692.

Small, Christopher. and Daniel Sousa. 2016. "Humans on Earth: Global Extents of Anthropogenic Land Cover from Remote Sensing." *Anthropocene, 14:* 1 – 33.

Taubenbock, Hannes, Thomas Esch, Andreas Felbier, Michael Wiesner, Achim Roth, and Stefan Dech. 2012. "Monitoring urbanization in mega cities from space." *Remote Sensing of the Environment*, *117:* 162-176.