DATA SCIENCE JOURNAL

Frye, C, et al. 2018. Using Classified and Unclassified Land Cover Data to Estimate the Footprint of Human Settlement. *Data Science Journal,* 17: 20, pp. 1–12, DOI: https://doi.org/10.5334/dsj-2018-020

RESEARCH PAPER

Using Classified and Unclassified Land Cover Data to Estimate the Footprint of Human Settlement

Charlie Frye, Earl Nordstrand, Dawn J. Wright, Carmelle Terborgh and Jeanne Foust

Environmental Systems Research Institute, Inc., US

Corresponding author: Charlie Frye (cfrye@esri.com)

Accurate, up-to-date maps of and georeferenced data about human population distribution are essential for meeting the United Nations Sustainable Development Goals progress measures, for supporting real-time crisis mapping and response efforts, and for performing many demographic and economic analyses. In December 2014, Esri published the initial version of the World Population Estimate (WPE) image service to ArcGIS Online. The service represents a dasymetric footprint of human settlement at 250-meter resolution. It is global and contains an estimate of the 2013 population for each populated cell. In 2016 Esri published an additional image service representing the earth's population in 2015 at 162-meter resolution. Esri's WPE is produced by combining classified land cover data indicating predominantly built-up or agricultural locations with Landsat8 Panchromatic imagery, road intersections, and known populated places. The model detects where settlement is likely to exist beyond the areas classified as predominantly built up. The result is a global dasymetric raster surface of the footprint of settlement with a score of the likelihood of human settlement for each cell of the footprint. Population data are apportioned to this settlement likelihood surface by overlaying population counts in polygons representing census enumeration units or political units representing population surveys. This paper presents the method developed at Esri for producing the estimate of settlement likelihood.

Keywords: Settlement; Gridded Population; Landsat 8; Landscape Disturbance

1 Introduction

Accurate, up-to-date maps of and georeferenced data about human population distribution are essential for meeting the United Nations Sustainable Development Goals progress measures, for supporting real-time crisis mapping and response efforts, and for performing many demographic and economic analyses (e.g., Blumenstock 2016; Geldmann, Joppa & Burgess 2014; and Martin, Maris & Simberloff 2016). To aid in these efforts, Esri published the initial version of the World Population Estimate (WPE) image service to its ArcGIS Online portal in December 2014. The WPE is a global dasymetric estimate of the footprint of human settlement, where each populated raster cell in the footprint contains an estimate of the number of people living there. In the summer of 2016, Esri published an additional image service for the world's population in 2015 based on an improved methodology for estimating the footprint. The improved methodology is presented here. With the 2015 estimate, Esri also released image services of population density, settlement likelihood scores, and confidence scores for settlement likelihood. **Figure 1** shows the global extent of the footprint as a map of population density.

Esri's image services are a type of web service, making one or more raster datasets available to ArcGIS software users via the ArcGIS Online portal. The services can be accessed by many users simultaneously. These services provide access to the data such that they do not need to be copied and can be used directly as input to ArcGIS geoprocessing tools and models or ArcGIS web application program interfaces (APIs). Use of these services is free for all registered ArcGIS software users.

The 2013 estimate is provided at a raster resolution (cell size) of 250-meters globally, and the 2015 estimate is provided at 162-meter resolution. One of the improvements to the 2015 estimate was moving the production from the Web Mercator Auxiliary Sphere coordinate system to the World Geodetic System WGS 1984 geographic coordinate system. This also means the raster resolution is expressed as the width of a cell at the equator. **Figure 2** shows an area in central Spain, with Madrid at the center, to illustrate the geographic precision or level of detail 162-meter resolution raster data provides.

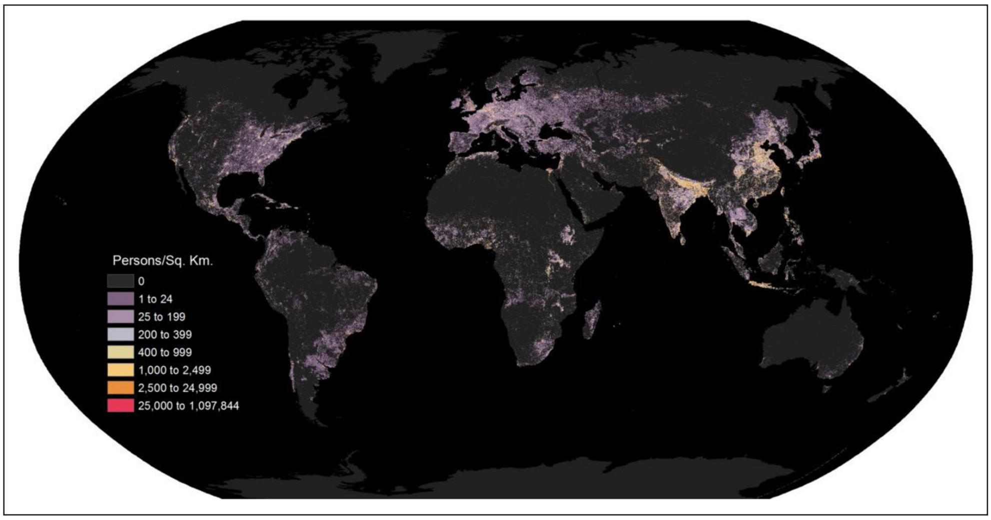


Figure 1: Global coverage of the WPE as a map of population density where raster cells on land represent a dasymetric surface. The populated cells are represented with an estimated density in units of persons per square kilometer.

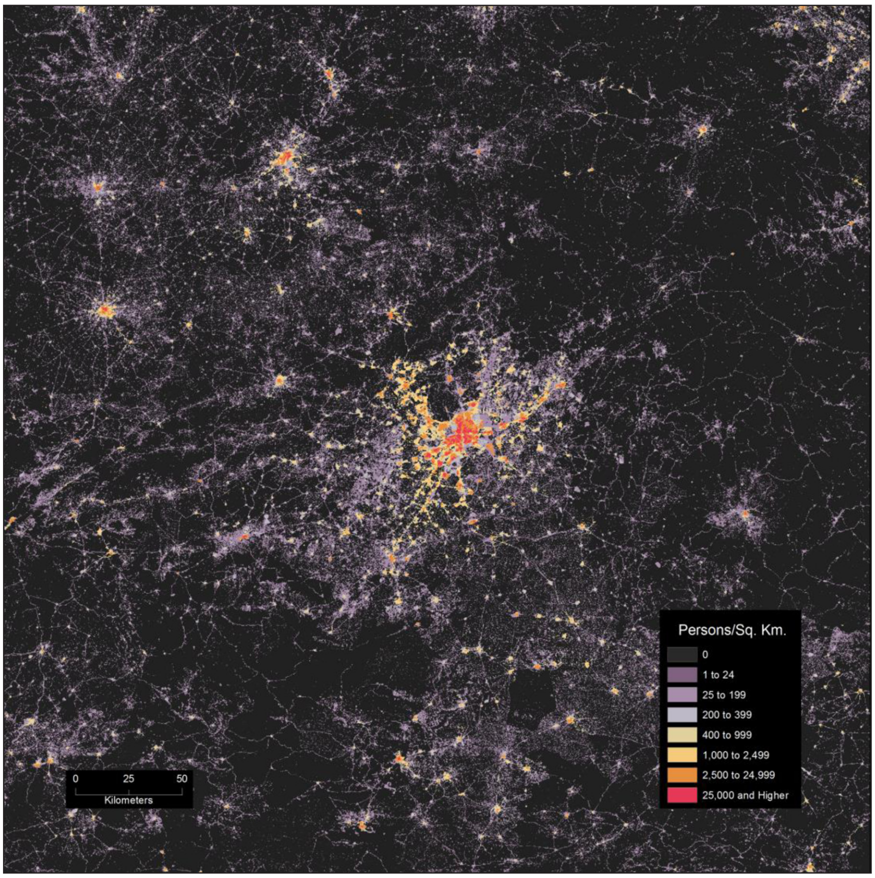


Figure 2: Madrid, Spain, and the surrounding territory representing the 162-meter raster resolution of the 2015 WPE.

Esri produced the WPE as a response to user's needs to perform economic and demographic analyses with a high level of geographic precision. The WPE is used by the Data Enrichment tool in Esri ArcGIS Online software and by Esri Business Analyst software. Other potential uses for the WPE include estimating populations affected by natural disasters, disease outbreak, and other humanitarian emergencies and estimating the impact of human activity on the landscape.

The additional services released in 2015, particularly the likelihood of settlement image service, expand the utility of the WPE by allowing Esri's users who possess census data or population estimates of higher quality than those used by Esri to produce higher-quality estimates of where people live. Some countries do not publicly release their most detailed census enumeration unit boundaries, but now ArcGIS users within those governments can produce higher-precision estimates of where their population may live by using the settlement likelihood service.

The WPE is a dasymetric portrayal, as opposed to an areal weighting or pycnophylactic interpolation, as described in Hay et al., (2005) that implements three smart interpolation methods for producing gridded population estimates. Linard and Tatem (2012) adjust this relationship by labeling smart interpolation as a fourth, more sophisticated approach involving imagery and multiple ancillary datasets. Thus, the WPE could be categorized as a dasymetric map using a smart interpolation method, though, as will be presented in the methodology section, the WPE is dasymetric first with smart interpolation applied afterward. Voinov (2014) presents arguments for the superiority of using a dasymetric technique to model the footprint of human settlement because assumptions about the distribution of population are made in the other methods.

The method for producing the WPE is different from other settlement pattern estimates in two ways. First, the unique combination of processing steps and the order of operations differ from other dasymetric settlement footprint estimates (Dobson et al., 2000; Pesaresi et al., 2012; Esch et al., 2013; European Statistical System, 2012; Linard and Tatem, 2012). Second, the model produces an estimate of the settlement footprint for nonurban areas using a technique to analyze panchromatic Landsat8 scenes to derive a measurement of likelihood of textural features (Haralick, Shanmugam, and Dinstein, 1973). Cheriyadat, et al., (2007) and Pesaresi, et al., (2012) implemented a textural features algorithm using a gray level co-occurrence matrix (GLCM), which is succinctly described in Albregtsen (2008) and Pradhan et al., (2013).

2 Methodology

Esri's model for producing the WPE begins with using a global 30-meter resolution classified land cover raster dataset called BaseVue 2013. The purpose is to establish known urban areas, areas likely to include settlement, and areas not likely to contain settlement. BaseVue is produced by MDA Information Systems LLC, a subsidiary of MacDonald Dettwiler and Associates Ltd (MDA, 2014). MDA derived BaseVue 2013 from Landsat8 scenes by using a classification and regression tree (CART) algorithm similar to Smith, Bolton, and Jengo (2004). BaseVue integrates the U.S. Geological Survey's 2011 National Land Cover Dataset (NLCD) (Homer et al., 2015) within the conterminous United States.

The BaseVue 2013 land cover dataset contains sixteen Anderson-style land cover classes (Anderson, et al., 1976). Esri's model begins by using the Reclassify tool from the ArcGIS Spatial Analyst software to assign an initial score to each BaseVue land cover class to represent the likelihood of people to live within a BaseVue cell with that respective class assignment. This includes a score of zero for cells with classes where people are not likely to live. **Table 1** lists BaseVue's classes and criteria (MDA, 2014) and **Table 2** lists scores assigned by Esri's model.

There are two challenges with this reclassified BaseVue data. First, there are missing populated places, particularly smaller places such as rural villages and farms that did not have sufficient levels of constructed materials per MDA's definitions for classes 20 and 21. Second, in the area with scores of 25, there is a great deal of unpopulated land. However, those areas also include populations living at the edges of cities or in rural areas, on farms, or in isolated residences.

To address the first issue of missing populated places, Esri used the GeoNames.org gazetteer database (GeoNames.org, 2013). Esri used a subset of GeoNames where the Category field contained a value of 'PPL', which GeoNames.org defines as 'a city, town, village, or other agglomeration of buildings where people live and work'. The following steps were taken using ArcGIS 10.3.1 software to process the GeoNames.org data and add their locations to the reclassified BaseVue dataset:

1. Buffer the GeoNames.org points by 100 meters to create a 200-meter diameter polygon for each point.
2. Add an integer score field to this buffers dataset and calculate all rows to a value of 150.
3. Convert the buffered polygons to a 30-meter resolution global raster dataset using the score attribute, which creates a raster with cell values of 150 or NoData.

Table 1: Land cover classes in BaseVue 2013. In particular, note the urban classes have a high requirement for infrastructure, which inherently excludes cells at the edges of urbanization or rural settlements such as farming villages.

Class	Class Name	Description
1	Deciduous Forest	Trees > 3 meters in height, canopy closure > 35% (<25% intermixture with evergreen species) that seasonally lose their leaves, except larch
2	Evergreen Forest	Trees > 3 meters in height, canopy closure > 35% (<25% intermixture with deciduous species), of species that do not lose leaves (will include coniferous larch regardless of deciduous nature)
3	Scrub/Shrub	Woody vegetation < 3 meters in height, > 10% ground cover. Only collect > 30% ground cover.
4	Grassland	Herbaceous grasses, > 10% cover, including pastureland. Only collect > 30% cover.
5	Barren or Minimal Vegetation	Land with minimal vegetation (<10%) including rock, sand, clay, beaches, quarries, strip mines, and gravel pits. Salt flats, playas, and non-tidal mud flats are also included when not inundated with water.
7	Agriculture, General	Cultivated cropland
8	Agriculture, Paddy	Cropland characterized by inundation for a substantial portion of the growing season
9	Wetland	Areas where the water table is at or near the surface for a substantial portion of the growing season, including herbaceous and woody species (except mangrove species)
10	Mangrove	Coastal (tropical wetlands) dominated by mangrove species
11	Water	All water bodies greater than 0.08 hectares (1 LS pixel) including oceans, lakes, ponds, rivers, and streams
12	Ice/Snow	Land areas covered permanently or nearly permanently with ice or snow
13	Clouds	Areas where no land cover interpretation is possible due to obstruction from clouds, cloud shadows, smoke, haze, or satellite malfunction
14	Woody Wetlands	Areas where forest or shrubland vegetation accounts for greater than 20% of vegetative cover and the soil or substrate periodically is saturated with or covered by water. Only used within the continental U.S.
15	Mixed Forest	Areas dominated by trees generally greater than 5 meters tall and greater than 20% of total vegetation cover. Neither deciduous nor evergreen species are greater than 75% of total tree cover. Only used within the continental U.S.
20	High Density Urban	Areas with over 70% of constructed materials that are a minimum of 60 meters wide (asphalt, concrete, buildings, etc.). Includes residential areas with a mixture of constructed materials and vegetation, where constructed materials account for > 60%. Commercial, industrial, and transportation, e.g., train stations, airports.
21	Medium-Low Density Urban	Areas with 30% to 70% of constructed materials that are a minimum of 60 meters wide (asphalt, concrete, buildings, etc.). Includes residential areas with a mixture of constructed materials and vegetation, where constructed materials account for greater than 40%. Commercial, industrial, and transportation, e.g., train stations, airports.

4. Use the Reclassify tool to set the NoData cells to 0 (zero). The Reclassify tool produces a new raster dataset where some or all of the values have been changed based on the user specifying input values that should be changed to new values.
5. Use the Local Statistics tool with the Maximum option, given the inputs of Step 4 and the reclassified BaseVue dataset. The Local Statistics tool performs statistical operations at the location of each cell for a series of overlapping raster datasets. Thus, the mean, minimum, or maximum value occurring at a given location may be derived. The result is a raster dataset representing all potentially populated areas.

The second problem of areas being included, particularly within agricultural areas where no people live, has a two-part solution. In the result of Step 5 above, any cell with a likelihood score of 25 needs to be evaluated to determine whether people are likely to be living there. Esri used two processes to accomplish this. The

Table 2: On the left are the BaseVue 2013 class identifiers and names, and on the right the initial remapping of BaseVue classes into settlement likelihood scores.

Class ID	Class Name	Modeled Population Likelihood Score and Rationale
1	Deciduous Forest	25—Potentially Orchard Agriculture
2	Evergreen Forest	0
3	Scrub/Shrub	0
4	Grassland	25—Potentially Range/Pasture Agricultural Land
5	Barren or Minimal Vegetation	0
7	Agriculture, General	25
8	Agriculture, Paddy	25
9	Wetland	0
10	Mangrove	0
11	Water	0
12	Ice/Snow	0
13	Clouds	0
14	Woody Wetlands	0
15	Mixed Forest	0
20	High Density Urban	200
21	Medium-Low Density Urban	150

first was to use proximity to road intersections. Tang (2003) argues that road pattern is closely related to urban growth. Thus, Esri created a global dataset of road intersections from two sources: HERE.com (HERE.com, 2015) and OpenStreetMap (OSMF, 2015). A country-by-country analysis was conducted to determine which countries had more data from either source and therefore which source would be used. The vector line street data were assembled into datasets for six continental zones. To produce intersection points, Esri used the following processing steps:

1. Create a geometric network for each zone's dataset. A by-product of creating a geometric network is a junction point feature class, representing all road intersection points. A geometric network is a geodatabase class that represents the connectivity of features from the input polyline feature classes.
2. Use the Append tool to put the six junction feature classes into one global point dataset. The Append tool concatenates features from one or more feature classes into a single output feature class that will contain all features.
3. Convert the junction dataset to a 30-meter resolution global raster dataset.
4. Use the Block Statistics tool with the Minimum option and a 5 × 5 cell neighborhood. Each cell is given a score of 150 if it is within the 5 × 5 cell of a cell corresponding to a junction point. The Block Statistics tool summarizes the values of the input raster into a coarser resolution output raster, allowing the user to derive means, minimums or maximums for a regular gridded set of "blocks" that conform to the origin of the input raster dataset.
5. Use the Con tool where the output of Step 5 above has a value of 25, and set it 150 if it corresponds with a cell value of 150 from the results of Step 4. The Con tool is used to apply conditional logic to the intersection of two raster datasets, and in particular allows complex algebraic statements as the expression of the logical relationship between cell values of the raster datasets at each cell's location.

To this point, the results contain scores of 150 or higher for areas that have a high certainty for representing places where people live. The cells with a score of 25 still represent mostly unpopulated locations but include farms and isolated residences. To screen out the areas where people are not living, Esri modeled panchromatic imagery from Landsat8 against the cells with a value of 25. The use of imagery for this purpose is common to all other country- or global-scale dasymetric settlement estimates (Cheriyadat et al., 2007; Pesaresi et al., 2012; Esch et al., 2013; European Statistical System, 2012; Linard and Tatem, 2012).

Esri's method of evaluating Landsat8 (Esri, 2014a) imagery assesses whether a 5 × 5 cell moving neighborhood window is likely to contain textural features (Haralick, Shanmugam & Dinstein, 1973; Albregtsen, 2008) as opposed to explicitly modeling the features as discussed by Cheriyadat et al., (2007) and Pesaresi, et al., (2012). Esri used the following processing steps in ArcGIS to produce a likelihood score for textural features to use as a likelihood of settlement score:

1. Create a 0.5 × 0.5-degree polygon grid in the WGS 1984 coordinate system. This grid is used as processing extents for the remainder of the model. The Fishnet tool was used to create this grid.
2. Extract the 15-meter resolution Landsat8 panchromatic imagery within the current processing grid extent. Note: A top of atmosphere (TOA) radiance correction was not available in the version of Esri's image service for Landsat8 panchromatic imagery. The Extract by Mask tool was used to perform this operation.
3. Create a cloud mask using values higher than 15,000. The values of the Landsat8 panchromatic imagery range from 5,000 (dark, or low reflectance) to 65,535 (white, or high reflectance). The Reclassify tool was used to assign cloud values to a value of 1 and non-cloud values to NoData.
4. Expand the cloud mask by 20 cells using the Expand tool. This was intended to remove incidental clouds or undetected clouds and cloud shadows that often occurred in cells adjacent to cloud cells.
5. Use the mask from Step 4 to assign NoData values to the result of Step 2. This removed the high-value cloud cells from the next steps of the analysis. The issue with including clouds is they create false positives for texture at their edges.
6. Use the Focal Statistics tool, with the results of Step 5 as input, to determine the range of cell values in a 5 × 5 cell rectangular neighborhood of each cell.
7. Use the Focal Statistics tool, with the results of Step 6 as input, to determine the sum of ranges of cells in a 5 × 5 cell rectangular neighborhood of each cell. This step eliminates the influence of single-cell spikes.
8. Determine the mean and standard deviation of values in the dataset (represents the processing grid's extent) resulting from Step 7. Add these to create a settlement texture threshold score.
9. Use the Raster Calculator tool with the Con function, with inputs of the result of Step 7, and the dasymetric surface such that the Con function selects values from the result of Step 7 above the settlement texture threshold score from Step 8, normalizes the range of those values from 1 to 100, and adds them to any non-zero score in the dasymetric surface. Cells with values below the threshold for settlement texture score in the dasymetric surface are set to 0 (zero).
10. Resample the result to 150-meter resolution using the nearest neighbor method. This resolution is 10 times the LandSat8 raster dataset and 5 times the land cover scores raster dataset.

Thus, the result of the above steps is a dasymetric surface representing settlement likelihood with values ranging from 0 to 300. **Figure 3** illustrates how these steps work on one example neighborhood.

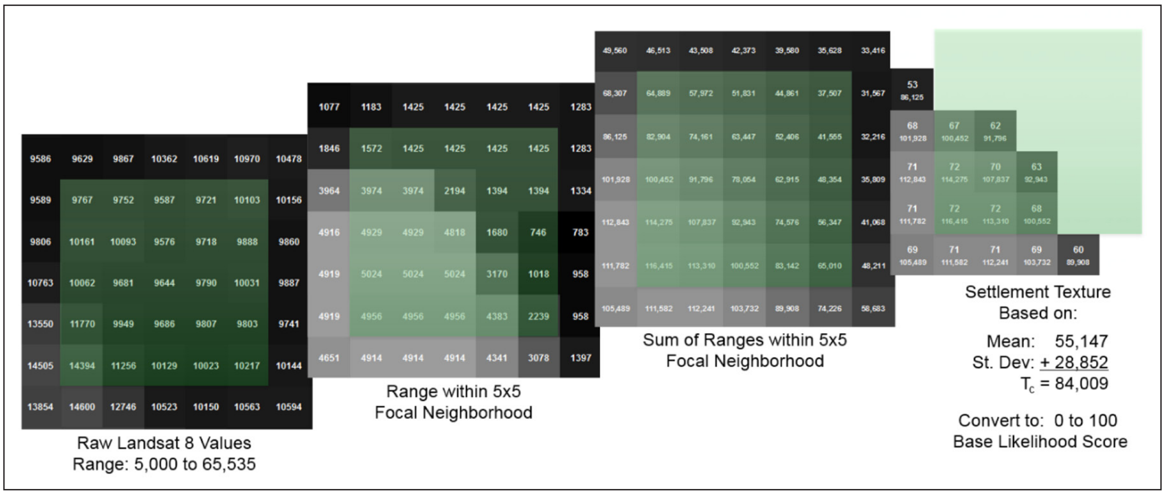
Any likelihood score above 150 represents at least low-medium residential population density. Any score above 200 represents high population density. Scores from 1 to 125 represent the potential for rural population density. Cells with a score of 0 represent no population living in that location. Statistically, the likelihood of settlement score can be represented as follows:

$$T_c = \sum_{i=c}^{25} \left(C_{Max} - C_{Min} \right)_{n = 5 \times 5}$$

Where T represents the likelihood of texture score within a 5 × 5 cell neighborhood of each cell (C) in the Landsat8 panchromatic imagery.

$$T_{Settlement} = Min\{T\mu_s + T\sigma_s > T_c\}$$

T-Settlement level is the likelihood of texture score for a cell most likely to indicate human settlement. The mean (μ) and standard deviation (σ) are for the 0.5-degree processing tile.



Raw Landsat 8 Values
Range: 5,000 to 65,535

9586	9629	9867	10362	10619	10970	10478
9589	9767	9752	9587	9721	10103	10156
9806	10161	10093	9576	9718	9888	9860
10763	10062	9681	9644	9790	10031	9887
13550	11770	9949	9686	9807	9803	9741
14505	14394	11256	10129	10023	10217	10144
13854	14600	12746	10523	10150	10563	10594

Range within 5x5
Focal Neighborhood

1077	1183	1425	1425	1425	1425	1283
1846	1572	1425	1425	1425	1425	1283
3964	3974	3974	2194	1394	1394	1334
4916	4929	4929	4818	1680	746	783
4919	5024	5024	5024	3170	1018	958
4919	4956	4956	4956	4383	2239	958
4651	4914	4914	4914	4341	3078	1397

Sum of Ranges within 5x5
Focal Neighborhood

49,560	46,513	43,508	42,373	39,580	35,628	33,416
68,307	64,889	57,972	51,831	44,861	37,507	31,567
86,125	82,904	74,161	63,447	52,406	41,555	32,216
101,928	100,452	91,796	78,054	62,915	48,354	35,809
112,843	114,275	107,837	92,943	74,576	56,347	41,068
111,782	116,415	113,310	100,552	83,142	65,010	48,211
105,489	111,582	112,241	103,732	89,908	74,226	58,683

53 86,125				
68 101,928	67 100,452	62 91,796		
71 112,843	72 114,275	70 107,837	63 92,943	
71 111,782	72 116,415	72 113,310	68 100,552	
69 105,489	71 111,582	71 112,241	69 103,732	60 89,908

Settlement Texture
Based on:

Mean: 55,147
St. Dev: + 28,852
T_c = 84,009

Convert to: 0 to 100
Base Likelihood Score

Figure 3: Main processing steps, starting with the Landsat8 panchromatic imagery; then range of values in a 5 × 5 cell neighborhood; then the sum of ranges; and finally, the cells with sums above the threshold for settlement texture score and their value once normalized.

Figure 3 also illustrates the process described in Steps 2–9 above, which produce the final dasymetric footprint of settlement scores representing a likelihood of settlement in those locations. To this point, the cells in the dasymetric settlement footprint represent the following conditions and value ranges:

- High Density Urban from BaseVue: 200 + 1 – 100 texture likelihood score
- Medium-Low Density Urban from BaseVue: 150 + 1 – 100 texture likelihood score
- Road Intersection or Place point buffers: 150 + 1 – 200 texture likelihood score
- Agricultural Lands from BaseVue: 25 + 1 – 100 texture likelihood score

The last part of Esri's model to create the WPE apportions population estimate data, stored in polygons representing either census enumeration units or areas where population has been estimated or surveyed. The sources for these data are as follows, in order of preference:

- Esri Demographics for United States and U.S. Protectorates and Territories: Polygon geographies are U.S. Census Block Group with Esri's current year estimate (Esri 2015b) with 218,125 polygon features.
- Canada: Environics Analytics at the Dissemination Area (DA) level with 56,204 polygons.
- Michael Bauer Research GmbH: 134 countries at admin level 3 (county) or 4 (city/town) for current year estimate, with 1,009,501 polygons.
- Seventy-eight countries with the most recent estimates from the United Nations Population Division, usually at admin level 3, though some are 2 (state) with 9,925 polygons.
- Eighteen countries with estimates more recent than the United Nations Population Division, totaling 316,368 polygons, though 315,826 are from Brazil.

The populations from these polygons are apportioned in a two-stage process to create a new raster dataset given the dasymetric settlement footprint with the likelihood scores as input. First, if there are more cells with nonzero settlement scores in a given polygon than the number of people, dasymetric settlement likelihood surface for that polygon is reprocessed as follows:

1. Reclassify the scores such that all scores of 123 and lower are set to 0 (zero).
2. Set new scores for the remaining cells by first subtracting 124 and then multiplying by 1.73.

Generally, the above two steps are required for sparsely populated polygons representing large geographic areas. To complete the apportionment, the populations of the polygons are distributed to the cells based on each cell's settlement likelihood score. This is done by determining the ratio of the polygon's population to the sum of settlement scores inside the polygon.

The resulting population surface is verified as having the correct total population for all countries and then for each country when the raster cells whose centers fall within a given country's boundary are extracted and summed. A population density surface is also computed, and locations with population densities over 50,000 persons per square kilometer are manually evaluated. Such areas occur due to horizontal inaccuracies within the population polygons, where they may have only cover or extend into locales where the dasymetric settlement likelihood surface contains too few cells. To address this, the topology of the population polygons is edited to match high-resolution satellite imagery (Esri 2015c), and the steps to apportion the population data are repeated.

3 Results and Discussion

To summarize, the method for determining the dasymetric settlement surface starts with known urban areas, adds highly likely locations near road intersections and GeoNames populated places, and then models the edges of settlement and fills in rural settlement using texture from panchromatic imagery. There are several aspects of this process worthy of further discussion:

- BaseVue urban area classifications
- Modeling the likelihood of textural features
- Areas of potential improvement

BaseVue urban classifications: The Medium-Low Density Urban and High Density Urban classifications require 40% and 60%, respectively, of land area to be commercial, industrial, or transportation, implying large areas of concrete exist. Thus, villages or concentrations of people where there are no paved roads or modern infrastructure are classified using the land cover type of the surrounding area. This also indicates a conservative estimate of the size of urbanized areas. The outskirts of cities and suburban areas may not contain a sufficient level of visible pavement or structures to be considered in either of the urban classes. It is important that these urban classes in BaseVue not overestimate the geographic extent of the urban footprint (Linard et al., 2012). Therefore, the urban classes were isolated and visually inspected relative to high-resolution satellite imagery in several dozen cities, occurring on all continents, to verify this to be true. Thus, these two classes of BaseVue were deemed an appropriate starting point.

Modeling the likelihood of textural features: Cheriyadat, et al. (2007) discuss an approach to using grayscale imagery as a basis for detecting dasymetric settlement footprints. Their approach focuses on finding local edge patterns (LEP) using a GLCM. Among the challenges they discuss is the lack of guidance for the size of the neighborhood to analyze around each pixel. Cheriyadat, et al. (2007) reported using a 17 × 17 neighborhood within a given image, and 31 × 31 at the edges.

As described above, a 5 × 5 neighborhood was used to analyze 15-meter resolution Landsat8 panchromatic imagery. This approximately matches Pesaresi, et al., (2012), where 5 × 5 cell neighborhoods were applied to 10-meter resolution imagery, though the basis for that determination was to use a 50-meter window size regardless of resolution. This neighborhood size was useful because it covers roughly half a city block and would encompass at least part of a large building and either its shadow or the shadow from an adjacent building. This combination of shadows and high reflectance off building materials is conceptually the same as LEP described above, except instead of defining explicit cell value and distance pairings, a high value range of cell values must occur within the 5 × 5 cell neighborhood.

Using only the range for a 5 × 5 cell neighborhood will provide some insight into the likelihood of textural features. However, it does not eliminate the possibility of one-cell spikes, particularly off highly reflective building materials or bodies of water. For example, such a spike would be a value of 48,000 in the midst of cell values ranging 7–11,000. The high range of values in the neighborhood of this cell would be skewed. However, using the sum of ranges sufficiently dampens the impact and reduces the likelihood for that location to be above the mean plus the standard deviation of the processing tile's sum of range values. This implies that two or more edges or textural features are within the 5 × 5 neighborhood of cells with a sum of value ranges higher than the mean plus the standard deviation for the processing tile. For comparison, Pesaresi, et al., (2012) used a minimum of four GLCM cell pairs occurring within a 5 × 5 neighborhood to constitute texture, thus showing that the count of cell pairs in a GLCM could also be used as a threshold.

Haralick, Shanmugam, and Dinstein (1973) presented textural features with the intent of locating and ultimately extracting features from a single image and presumed the values of grayness in the cells could be represented using pre-specified pairs of values, rather than ranges between any pair of cells. This makes sense if a feature can be fully defined using pairs of cells. The values of cells in the same location but having a different image date, or for the same type of structure (e.g., a McDonald's restaurant) but in a different image location, will vary even if all imagery is from the same sensor. Thus, a method that is more tolerant of varying conditions is needed.

This idea of finding a relatively high range of cell values within a local neighborhood is not new. Here it has been adapted from terrain ruggedness or rugosity models found in the landscape ecology literature. Rugosity is a measurement of the roughness of terrain (normally quantified by a ratio of surface area to planar area) and is useful for describing the geomorphic conditions that potentially define major characteristics for plant and animal habitat (Beier, Majka, and Spencer, 2008; Jenness, 2004). Models for ruggedness analyze elevation datasets and often a slope derivative to classify how much change exists within a locale (Cooley, 2016). This can be applied to grayscale imagery, whereby high amounts of local changes in value may indicate landscape disturbance, settlement, or false positives such as sun glints on water bodies or the edges of natural or agriculturally vegetated areas. In producing the WPE, many of these false positives were inconsequential because they did not occur within urban or agricultural areas, and others were eliminated later when the lower textural values in sparsely populated areas were removed.

Areas to improve: This method could be improved in several areas:

- Horizontal accuracy of the census polygon data: Accuracy of boundary data are improving; however, more improvements are needed. Balk, Yentman, and de Sherbinin (2010) note this issue, and Bhaduri, et al., (2007) discuss the impact of spatially inaccurate data on data integration workflows. Just over 4.8% of the populated cells in the WPE were at a population density of over 25,000 persons per square kilometer. This percentage seems extremely high. Horizontal inaccuracy of vector data can result in forcing large populations into a small number of cells.
- Need better census data for many countries: The population estimates for twenty-two countries were ten or more years old.
- Land use data: The U.S. Census data boundaries for block groups contain many polygons with no population. These correspond to industrial plants, factories, airports, and other facilities where no people reside. Similar data are needed for most other countries.
- Modifiable Areal Unit Problem (MAUP) (Openshaw, 1984): MAUP varies by scale in this work due to the arbitrary choice of raster resolution. For convenience, the 30-meter resolution MDA BaseVue 2013 land cover data, which are derived from Landsat8, are used as a snap raster dataset for much of the processing. However, if the processing were to be done at 90-meter or 150-meter resolution, some variation in the footprint pattern could result.
- Use of 0.5-degree processing tiles: The 0.5-degree processing tiles provide the basis for the mean and standard deviation of the sum of ranges for the Landsat8 panchromatic data. This is arbitrary and, ultimately, an unnecessary artifact of an ArcGIS geoprocessing workflow. Future workflows are planned that will eliminate this aspect of the workflow. However, a tiled workflow does provide a very conservative basis for the regional mean and standard deviation of the sum of ranges. Newer technology exists in the form of mosaic datasets, which can process the entire globe at 15- or 30-meter resolution in a single step. Therefore, another basis will be needed to derive the regional mean and standard deviation. Jones and O'Neill (2013) used a 100-kilometer neighborhood for this purpose.
- Artifacts from the processing grid and Landsat8 scene edges: Because the Landsat8 scenes do not have equivalent distributions or identical ranges of cell values, the 0.5-degree processing grid used to produce the dasymetric footprint of settlement intersects the Landsat8 scenes in a spatially independent way. Statistics for processing tiles containing only a small percentage of cells on land and a smaller percentage representing settlement may be skewed. **Figure 4** illustrates an extreme case of the impact of these limitations. There are also areas in the final output showing artifacts along the edges of Landsat8 scenes. This is due to two factors: adjacent scenes that are relatively cloud free may be more than two months apart, and at the time of processing, the information necessary to include a TOA correction for radiance was not yet available. The latter would help normalize the differences in distribution and cell value range between adjacent scenes. The former could be addressed by targeting leaf-off imagery whenever possible.

One aspect of this method is not starting with or using any previously produced gridded population estimate, gridded settlement footprint estimate, or lights at night estimate. This affords the benefit of not including any of the potential errors or uncertainty of those works. However, each of these previous efforts also have strengths, and therefore this method derives no benefit from the existing gridded population datasets. It

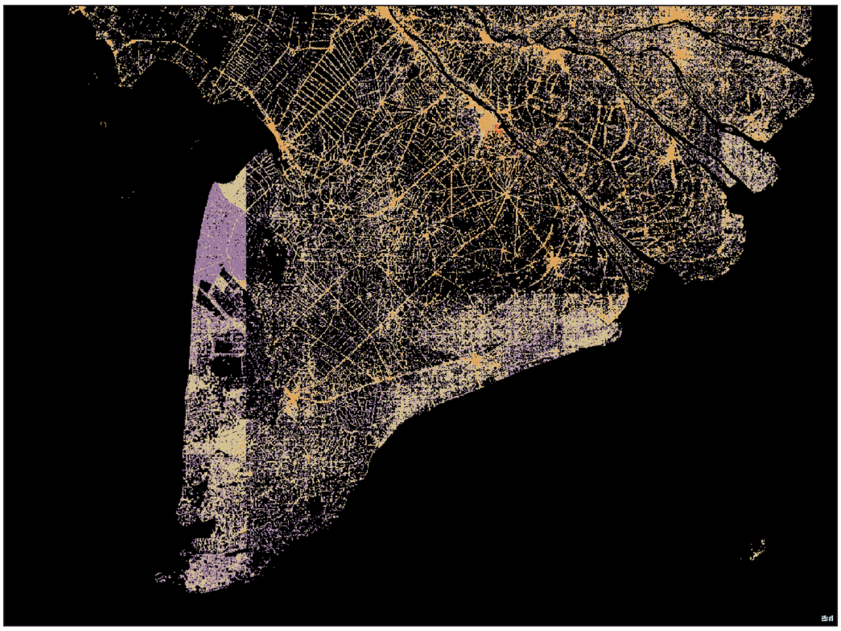


Figure 4: On the southern coast of Cambodia, a clear indication of the extents of the 0.5 degree processing grid can be seen.

should also be noted that, as made apparent in the above section of areas to improve, many of the ancillary input datasets vary in spatial quality, and therefore, any efforts to model a dasymetric settlement surface using these datasets suffers from the uncertainty of where the quality varies.

4 Conclusion

The authors propose that the methodology described herein for determining a dasymetric settlement likelihood surface, starting with a conservative estimate of urban land cover and then adding to the footprint based on known locations of settlement, locations near road intersections, and the likelihood of two or more textural features within a 5 × 5 cell neighborhood of Landsat8 panchromatic imagery, is not only a viable alternative to the previously published smart interpolation methods for producing a dasymetric settlement footprint but may also be considerably faster. In particular, using a GLCM to confirm the presence of edges of textural features appears to require a great deal of computational resources. These resources are needed to test for specific conditions within potentially large spatial neighborhoods as opposed to the minimalist 5 × 5 cell neighborhood undertaken in this method.

Acknowledgements

MDA Information Systems LLC – BaseVue classified land cover data
U.S. Geological Survey – Landsat8 Data
Amazon AWS – S3 volume access for Landsat8 panchromatic data, and Landsat8 metadata
here.com – Streets and roads data
OpenStreetMap.org – Streets and roads data
GeoNames.org – Populated places data
Michael Bauer Research GmbH – Population estimates
Environics Analytics – Population estimates
United Nations Population Division – Population estimates
Esri – Population estimates

Competing Interests

The authors have no competing interests to declare.

References

Albregtsen, F. 2008. Statistical Texture Measures Computed from Gray Level Coocurrence Matrices, Unpublished. Available at: http://www.uio.no/studier/emner/matnat/ifi/INF4300/h08/undervisningsmateriale/glcm.pdf [Last accessed 8 October 2016].

Anderson, JR, Hardy, EE, Roach, JT and **Witmer, RE.** 1976. A Land Use and Land Cover Classification System for Use with Remote Sensor Data. *U.S. Geological Survey Professional Paper 964, A revision of the land use classification system as presented in U.S. Geological Survey Circular 67*. United States Government Printing Office, Washington, DC.

Balk, D, Yentman, G and **de Sherbinin, A.** 2010. Construction of Gridded Population and Poverty Data sets from Different Data Sources. *E-Proceedings of European Forum for Geostatistics Conference*. Tallinn, Estonia.

Beier, P, Majka, DR and **Spencer, WD.** 2008. Forks in the Road: Choices in Procedures for Designing Wildland Linkages. *Conservation Biology*, 22(4): 836–851. DOI: https://doi.org/10.1111/j.1523-1739.2008.00942.x

Bhaduri, B, Bright, E, Coleman, P and **Urban, ML.** 2007. LandScan USA: A High-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal*, 69(1–2): 103–17. DOI: https://doi.org/10.1007/s10708-007-9105-9

Blumenstock, JE. 2016. Fighting Poverty with Data. *Science*, 353: 753. DOI: https://doi.org/10.1126/science.aah5217

Cheriyadat, A, Bright, E, Potere, D and **Budhendra, B.** 2007. Mapping of settlements in high-resolution satellite imagery using high performance computing. *GeoJournal*, 69(1–2): 119–29.

Cooley, S. 2016. Terrain Roughness – 13 Ways. Available at: http://gis4geomorphology.com/roughness-topographic-position/ [Last accessed 8 October 2016].

Dobson, JE, Bright, EA, Coleman, PR, Durfee, RC and **Worley, BA.** 2000. Landscan: A Global Population Database for Estimating Populations at Risk. *Photogrammetric Engineering & Remote Sensing,* 66(7).

Esch, T, Marconcini, M, Felbier, A, Roth, A, Heldens, W, Huber, M, Schwinger, M, Taubenbock, H, Muller, AA and **Dech, S.** 2013. Urban Footprint Processor—Fully Automated Processing Chain Generating Settlement Masks from Global Data of the TanDEM-X Mission. *IEEE Geoscience and Remote Sensing Letters*, 10(6): 1617–621. DOI: https://doi.org/10.1109/LGRS.2013.2272953

Esri. 2014a. Current Landsat8 Image Services in ArcGIS Online. *ArcUser*, Spring 2014.

Esri. 2015b. Methodology Statement: 2015/2020 Esri US Demographics Updates. *An Esri White Paper*, March 2015. Available at: http://downloads.esri.com/esri_content_doc/dbl/us/J10268_Methodology_Statement_2015-2020_Esri_US_Demographic_Updates.pdf [Last accessed 14 October 2016].

Esri. 2015c. World Imagery Map Image Layer. Available at: https://www.arcgis.com/home/item.html?id=10df2279f9684e4a9f6a7f08febac2a9 [Last accessed March–May, 2015].

European Statistical System, ESSnet project GEOSTAT. 2012. GEOSTAT 1A – Representing Census Data in a European Population Grid. *European Forum for Geostatistics*. Available at: http://ec.europa.eu/eurostat/statistics-explained/index.php/Population_grids [Last accessed 8 October 2016].

Geldman, J, Joppa, LN and **Burgess, ND.** 2014. Mapping Change in Human Pressure Globally on Land and within Protected Areas. *Conservation Biology*, 28: 1604–1616. DOI: https://doi.org/10.1111/cobi.12332

GeoNames.org. 2013. GeoNames. Available at: http://geonames.org/export/dump [Last accessed June 2013].

Haralick, RM, Shanmugam, K and **Dinstein, I.** 1973. Textural Features for Image Classification. *IEEE Transactions on Systems, Man and Cybernetics*, Vol. SMC-3(6): 610–621, Nov 1973. DOI: https://doi.org/10.1109/TSMC.1973.4309314

Hay, SI, Noor, AM, Nelson, A and **Tatem, AJ.** 2005. The Accuracy of Human Population Maps for Public Health Application. *Trop Med Int Health*, 10(10): 1073–1086. DOI: https://doi.org/10.1111/j.1365-3156.2005.01487.x

Here.com. 2015. HERE Map Data. Available at: https://here.com/en/products-services/data/here-map-data.

Homer, C, Dewitz, J, Yang, L, Jin, S, Danielson, P, Zian, G, Coulston, J, Herold, N, Wickham, J and **Megown, K.** 2015. Completion of the 2011 National Land Cover Database for the Conterminous United States – Representing a Decade of Land Cover Change Information. *Photogrammetric Engineering & Remote Sensing*, 81(5).

Jenness, JS. 2004. Calculating landscape surface area from digital elevation models. *Wildlife Society Bulletin*, 32: 829–839. DOI: https://doi.org/10.2193/0091-7648(2004)032[0829:CLSAFD]2.0.CO;2

Jones, B and **O'Neill, BC.** 2013. Historically Grounded Spatial Population Projections for the Continental United States. *Environmental Research Letters*, 8. DOI: https://doi.org/10.1088/1748-9326/8/4/044021

Linard, C, Gilbert, M, Snow, RW, Noor, AM and **Tatem, AJ.** 2012. Population Distribution, Settlement Patterns and Accessibility across Africa in 2010. *PLoS ONE*, 7(2): e31743. DOI: https://doi.org/10.1371/journal.pone.0031743

Linard, C and **Tatem, AJ.** 2012. Large-scale spatial population databases in infectious disease research. *International Journal of Health Geographics*, 11(1): 7. DOI: https://doi.org/10.1186/1476-072X-11-7

MacDonald, Dettwiler and Associates Ltd. (MDA). 2014. BaseVue 2013. Available at: http://www.arcgis.com/home/item.html?id=1770449f11df418db482a14df4ac26eb [Last accessed 14 October 2016].

Martin, JL, Maris, V and **Simberloff, DS.** 2016. The need to respect nature and its limits challenges society and conservation science. *Proceedings of the National Academy of Sciences*, 31(2): 6105–6112. DOI: https://doi.org/10.1073/pnas.1525003113

Openshaw, S. 1984. *The Modifiable Areal Unit Problem*. Geobooks, Norwich, England.

OpenStreetmap Foundation (OSMF). 2015. *OpenStreetMap*. Available at: https://www.openstreetmap.org.

Pesaresi, M, Blaes, X, Ehrlich, D, Ferri, S, Gueguen, L, Haag, F, Halkia, M, Heinzel, J, Kauffmann, M, Kemper, T, Ouzounis, GK, Scavazzon, M, Soille, P, Syrris, V and **Zanchetta, L.** 2012. A Global Human Settlement Layer from Optical High Resolution Imagery. *JRC Publications Repository*. Publications Office of the European Union. Available at: http://publications.jrc.ec.europa.eu/repository/handle/JRC77925 [Last accessed 14 October 2016].

Pradhan, B, Hagemann, U, Tehrany, MS and **Prechtel, N.** 2013. An Easy to Use ArcMap Based Texture Analysis Program for Extraction of Flooded Areas from TerraSAR-X Satellite Image. *Computers and Geosciences,* 63(Feb 2014): 34–43. DOI: https://doi.org/10.1016/j.cageo.2013.10.011

Smith, FGF, Bolton, C and **Jengo, C.** 2004. The Classification of Hyperspectral Data using the CART Classification Approach. *Proceedings of the ASPRS 2004 Annual Meeting*. Denver, CO, USA.

Tang, J. 2003. Evaluating the Relationship between Urban Road Pattern and Population Using Fractal Geometry. UCGIS.org. Available at: http://www.ucgis.org/summer03/studentpapers/junmeitang.pdf [Last accessed 14 October 2016].

Voinov, S. 2014. Modeling Population Distribution Based on EO-Derived Data on the Built-Environment. Master's Thesis in Photogrammetry and Geoinformatics, Stuttgart University of Applied Sciences. Available at: http://elib.dlr.de/97361/1/PG3_Voinov_Thesis.pdf [Last accessed 14 October 2016].

How to cite this article: Frye, C, Nordstrand, E, Wright, DJ, Terborgh, C and Foust, J. 2018. Using Classified and Unclassified Land Cover Data to Estimate the Footprint of Human Settlement. *Data Science Journal,* 17: 20, pp. 1–12, DOI: https://doi.org/10.5334/dsj-2018-020

Submitted: 31 October 2016 **Accepted:** 28 August 2018 **Published:** 06 September 2018

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press. **OPEN ACCESS**