Documentation for the World Population Estimated (WPE) Data Set

November 2016

Environmental Systems Research Institute

Abstract

The World Population Estimated data by Esri are global dasymetric ~150-meter resolution raster datasets representing where people live. There are four datasets:

- Estimated Population
- Estimated Population Density (1-km resolution)
- Estimated Settlement Likelihood
- Confidence Surface

The WPE uses MDA's BaseVue 2013, which is a 30-meter resolution Landsat8-based classified land cover product which contains relevant classes for medium and high density developed areas, and classes of vegetated land cover where people likely may be living. Esri uses those latter classes as a processing mask. Within that mask the presence of road intersections (150-m resolution) the locations for GeoNames populated places (1-km buffers), and Landsat8 Panchromatic texture were combined to determine the foot print and relative likelihood of low-density settlement.

To assign population to the estimated settlement likelihood surface, polygon data representing the finest level of census or surveyed population estimates were collected and used to apportion a population estimate onto each cell of the settlement likelihood surface. In the 2015 edition of the WPE, 1.61 million polygons were used.

All the GIS processing to create the WPE is done using the WGS_1984 geographic coordinate system, in part because of the 1:1 cell matching with the UTM zone Landsat8 data.

To produce the population density estimate, a 0.5-degree cell size WGS_1984 raster was created and given attributes for the percentage of area each row of cells represented of the area of the cells at the equator. This dataset could be used in a raster calculator expression with the population estimate to provide the estimated density.

The confidence surface reflects two dimensions that affect confidence. First is the ratio of the population polygon's area to the count of people estimated to live within the area represented by the polygon. Second is the effect of raster GIS processing, in the form of the impact of resampling (Nearest Neighbor was the best performing resampling method). Resampling potentially introduces error in dasymetric raster surfaces where of populated footprint cells abut zero-population cells. This occurs at coastlines and edges of population centers, and the more complex the edge, the higher the prospect for error. Esri's estimate is a 2.5 to 4.0% margin of error is introduced given the current raster

processing resolutions. Confidence is expressed on a 1 to 5 scale, with 1 being the highest level of confidence. Future versions will also incorporate an estimate of the quality of the census based on collection method and age of the census.

Data set citation:

Esri. 2015. World Population Estimated. https://doi.org/10.13140/RG.2.2.16160.79367

The data are available as GIS Web Services as follows:

2013 Edition

http://www.arcgis.com/home/item.html?id=f42926b976064414ac0d03ae51aa4f2e
https://landscape6.arcgis.com/arcgis/rest/services/World_Population_Estimated/ImageServer

2015 Edition

http://univredlands.maps.arcgis.com/home/item.html?id=2417dafad0b54276a2333d76a0b27311
http://landscape7.arcgis.com/arcgis/rest/services/World_Population_Estimated_2015/ImageServer

http://www.arcgis.com/home/item.html?id=625e9da1afed40b78aaf412f519b22d3
http://landscape7.arcgis.com/arcgis/rest/services/World_Population_Estimated_Density_2015/ImageServer

http://landscape7.arcgis.com/arcgis/rest/services/World_Population_Estimated_Probability_2015/ImageServer
http://landscape7.arcgis.com/arcgis/rest/services/World_Population_Estimated_Confidence_2015/ImageServer

Suggested citation for this document:

Frye, C., and Nordstrand, E., Environmental Systems Research Institute (Esri). 2017. Documentation for the World Population Estimated (WPE) Data Set. 380 New York Street, Redlands, CA 92373.

We appreciate feedback regarding these data such as suggestions, discovery of errors, difficulties in using the data, and format preferences. Please contact:

Environmental Systems Research Institute, Inc.
c/o Charlie Frye
380 NewYork Street
Redlands, CA 92373
Phone: 1 (909) 793-2853
Email: cfrye@esri.com

Contents

I. Introduction

In December 2014, Esri published the initial version of the World Population Estimate (WPE) image service to ArcGIS Online. The service represents a footprint of human settlement at 250-meter resolution, is global, and contains an estimate of the 2013 population for each populated cell. Figure 1 illustrates the coverage. A Poster-size PDF derived from Figure 1 is available online. This means raster cells on land were either represented as unpopulated or with an estimated population count. In the summer of 2016 Esri published an update representing the estimated footprint and population for 2015. The 2015 estimate included additional image services for population density and confidence. The 2015 services were all based on 162-meter resolution (see Figure 2), which given the following note we often describe as being 150-meter resolution, which is closer to the average cell size. (Note: this resolution is for a raster dataset using the World Geodetic System WGS 1984 geographic coordinate system, which means this is the size of a cell at the equator, and the farther a cell is from the equator, the smaller the area of the earth it represents.) Esri is currently processing data for a 2016 update, which is planned to be published later this year.

This white paper is intended to introduce the nature and purpose of the WPE services, provide recommended steps to get started using the services, and to specify essential details about the methodology used to produce the WPE.

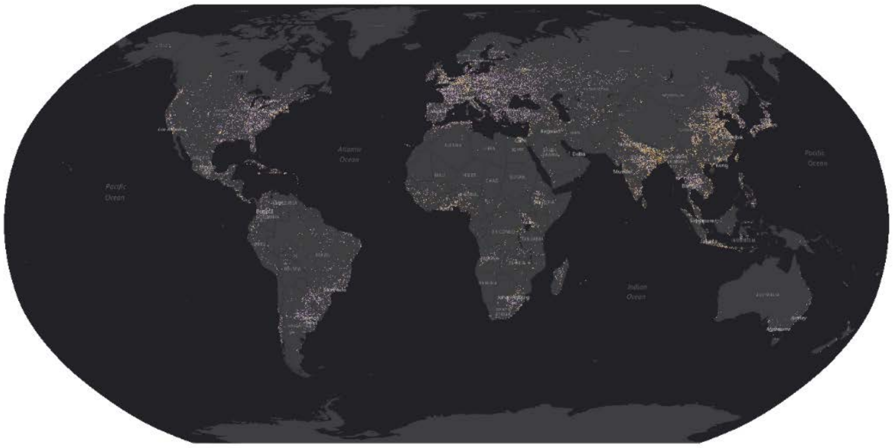


Figure 1. WPE global coverage

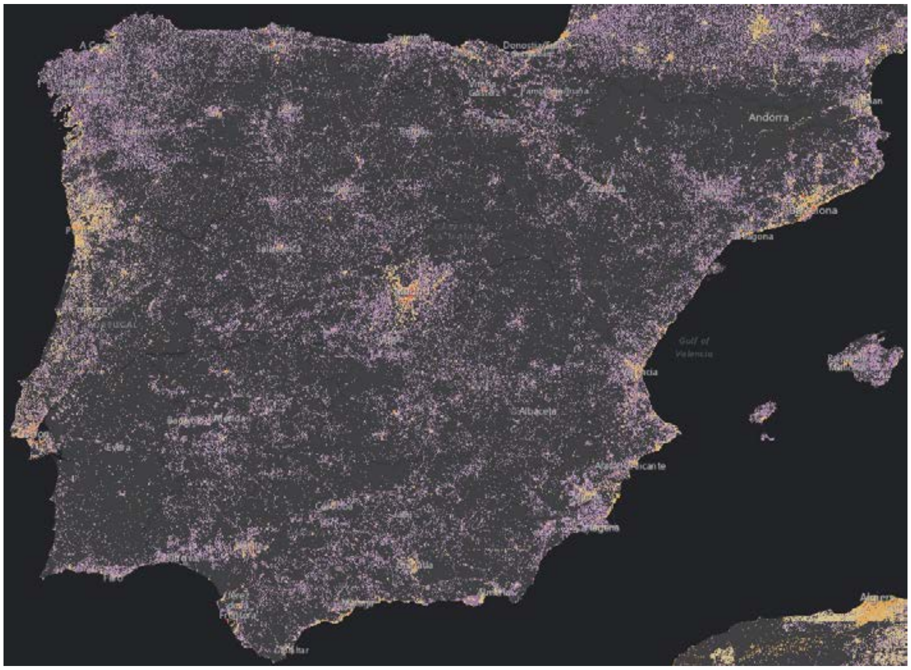


Figure 2. Iberian Peninsula example of WPE Density service used in an online map and illustrating the level of detail available.

A. Overview

The WPE is a dasymetric raster surface where each cell represents an estimated count of people living in the location the cell represents. The same method is applied globally, which means the values in each of the cells are intended to be comparable. Dasymetric means the production method specifically accounts for areas where people do not live. Esri used classified land cover data to specifically exclude areas where people do not live, such as atop high mountains, on water bodies, or on mountains.

By “living in the location”, we mean the WPE is intended to show where people live most of the time. This corresponds to the idea of a “census de jour”, which enumerates where individuals usually reside, regardless of where they are located on the day the census is taken.

This white paper is intended to introduce the nature and purpose of the WPE services, recommended steps to get started using the services, and to provide essential details about the methodology used to produce the WPE.

At a high level, the WPE is produced by first identifying areas that are well agreed or likely to be where people live and do not live. Then texture within Landsat8 imagery consistent with human settlement is added This produces a discreet footprint of settlement. Then population data, in the form of enumeration unit polygons, are apportioned onto the footprint. The population data is sourced from censuses of various countries, and from commercial data and United Nations data with population estimates as surrogates for censuses. The complete list of sources will be covered later in this paper. These main sources represent one of the challenges the WPE is designed to overcome. That challenge is to transform even the lowest quality census data, which can only be said to represent a survey rather than a formal de jure census, and which might be quite old by modern census taking standards, and have coarse (large) enumeration units lacking geographic specificity, into a geographically precise estimate of where people live.

Key to the success of this process is how Esri mitigates false positive and negative aspects of the footprint because satellite imagery alone cannot determine whether people live in certain structures versus others. Confounding factors such as clouds and high levels of reflectance off natural features such as surface water are statistically similar to reflectance off human made structures and the adjacent contrast from shadows cast by these structures. Thus, additional data, including classified land cover, road intersections, and named place points are used to reinforce the footprint with known locations of population or areas where people are not likely to live.

B. Purpose for Producing the WPE

Esri's purpose in doing so is to facilitate population-based spatial analysis in ArcGIS software, specifically, providing a footprint for ArcGIS tools to:

- Estimate demographic characteristics with higher geographic fidelity than traditional choropleth mapping techniques.
- Estimate where economic behaviors, political attitudes, and cultural characteristics occur with higher geographic precision.
- Estimate location and count of populations affected by:
    - Natural Disasters and Complex Humanitarian Emergencies (for mitigation, response, & recovery)
    - Exposure to disease (to prevent transmission)
    - Disease outbreak (to direct effective treatment efforts)
- Estimate human impact on the environment for purposes of sustainability, resilience, and planning along the lines of green infrastructure.

C. Criteria for Success

In certifying the WPE datasets for release as web services there are several criteria used to ensure the quality. There are as follows:

i. The total number of people

The raster dataset that is the basis for our web service must have an estimated of population within 1.000% of the U.N. Statistical Division's current year estimate. The variance is due to different sources being used. The WPE estimate for 2015 is 7,287,982,892 and the U.N. estimate is 7,324,782,000. The difference is 0.512%.

To check the population represented by a raster dataset, we add a field to the attribute table and calculate it to be the value times the cell count for that value, and the sum of values in the new field is equal to the total population. Pay special attention to the discussion in Sections II.A and II.B about the effect of changing the projection of the raster, even by just specifying a different output coordinate system for the image service. Those sections detail how to properly avoid misusing the services.

ii. The proportion of people in major density categories

That is to say, we determine what percentage of people are likely to live in urban densities versus rural densities. Because definitions for urban vary, and include more than only density, our certification is approximate, but definitively within range of expectations. At the lower end, we use the United Nations estimate of 54% and the World Bank's estimate of 53.9% of people living within urban areas in 2015. On the high end, 80.7% represents the United States level, which is based on a generous density threshold of just under 500 people per square kilometer, as opposed to another commonly used threshold of 2,500 persons per square kilometer.

In the 2013 WPE, there were approximately 55.47% of people estimated to be living in an urban density.

In the 2015 WPE, we additionally produced a density dataset, which made our estimate more precise, we estimated 46.93% lived at a density of 2,500 persons per square kilometer or higher, and 72.96% of people lived in a density of 500 persons per square kilometer or higher.

II. Recommended Steps to Get Started

The WPE is intended to be used online for mapping and visualization. Perhaps the most obvious use case is including the population density layer in lieu of a choropleth map of population density, which effectively declutters the resulting map of extraneous information, and allows population to contextualize other information rather than visually dominating the map.

The WPE is also intended to support mapping, visualization, and spatial analysis in Desktop GIS workflows. This section is particularly dedicated to the purpose of creating awareness for the aspects of the WPE that affect spatial analysis. The fact that the image services are based on raster datasets representing spatially varying discreteness of different statistical data types means specific knowledge and precautions must be used in order to ensure minimal distortions and data loss occur during spatial analysis workflows. The traditional rules of thumb for processing raster data in spatial analysis workflows, e.g., relating to resampling, projections, and cell size are not sufficient to ensure success when working with the WPE.

The following sections present essential understanding to successfully use the WPE layers in spatial analysis workflows.

A. Make Image Server Layer Tool

Most of the image services hosted by Esri are set, by default, to use the Web Mercator projected coordinate system. This is done to ensure the default online map viewing experience has the fastest possible draw times. However, the Web Mercator projected coordinate system is not appropriate to use when performing spatial analysis because it distorts distance and area measurements, causing erroneous outcomes.

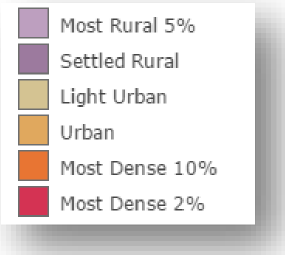
ArcGIS Desktop applications allow users to bypass using the Web Mercator projected coordinate system altogether when using image services. Esri has produced a blog entry explaining the details for how to do so, called, “Use Living Atlas Image Services in Your Desktop Analysis”. In particular, the workflow described overcomes the default behavior of ArcGIS Online, which adds the layer to ArcGIS desktops already set in the Web Mercator Projected Coordinate System.

B. Considerations for Projecting WPE (any raster) Data

The most important information to know relative to the workflow referenced in the previous paragraph is that the raster datasets used by the image services are stored using the WGS 1984 geographic coordinate system. This means two options are available. The first is to use the above workflow and specify you want to receive the data in WGS 1984, which will result in the service supplying data without transforming its coordinate system. The second option is to specify the resulting layer in an analytically useful coordinate system. We recommend Mollweide or Equal Area Cylindrical for analyses that depend on areal comparisons. We recommend Equidistant Cylindrical for analyses that require distance. Projecting from the source WGS 1984 geographic coordinate system to any of these projected coordinate systems needed for analysis will result in some data loss, but these recommendations have been tested to ensure the minimum amount of data loss when projected from WGS 1984. We also tested projecting the full global WGS 1984 dataset to Web Mercator, which changed the total population from 7,126,379,999 to 8,583,012,070. That’s an increase of just over 20%. Then we tried projecting the result back to WGS 1984, and only got 6,378,036,269 people.

C. About the colors used to portray the WPE services

The 2015 WPE data are presented in using the legend shown in Figure 3:



🟪 Most Rural 5%
🟪 Settled Rural
🟨 Light Urban
🟧 Urban
🟧 Most Dense 10%
🟥 Most Dense 2%

Figure 3. World Population Estimated Density Layer’s Legend

The cut points for these population density classes are:

- Most Rural 5%: 100 persons/sqkm or less
- Settled Rural: 400 persons/sqkm
- Light Urban: 1,500 persons/sqkm
- Urban: 2,500 persons/sqkm
- Most Dense 10%: 16,979 persons/sqkm
- Most Dense 2%: 29,000 persons/sqkm or more

Figure 4 illustrates another way to visualize the population using these colors, where the distribution of the world's population is displayed as it radiates outward from the center starting at 1 person per square kilometer and turning clockwise toward the highest densities. Note the 200 threshold for rural is shown in lieu of the most rural 5% threshold used in the web services.

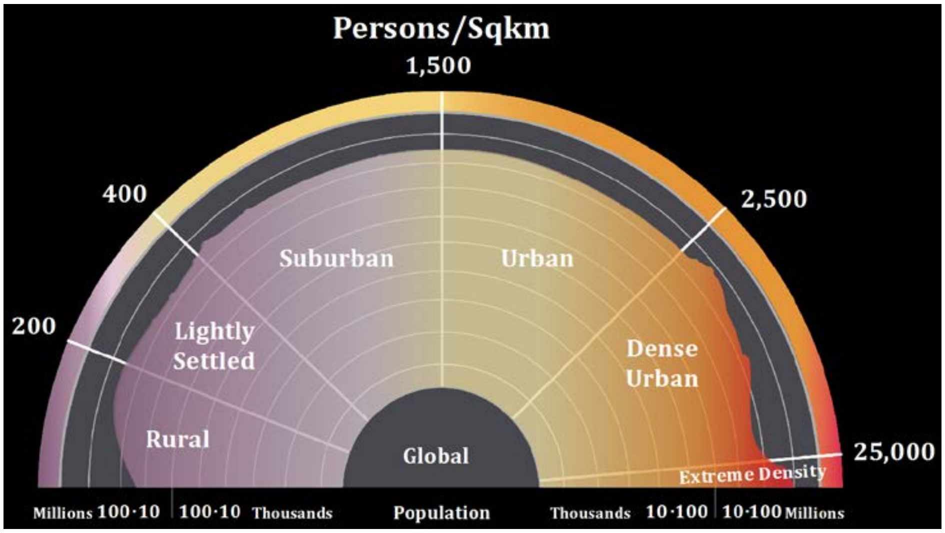


Figure 4. World's population by estimated density graph.

D. Special Nature of Highly Discrete Count Raster Data

Given the figures in the above example, Esri did some additional tests, including varying other parameters pertaining to the output. The results were surprising indicating the cell size and resampling method parameters, which are set in ArcGIS Desktop's geoprocessing environment settings, had a dramatic impact. To understand the impact of varying these parameters, Esri evaluated the total cell count, the NoData cell count, the total population, the range of cell values by checking min and max values. To verify the total population, we add a new field to the raster dataset's attribute table, and calculate that field to be the product of the value field and the count field. *(**Note**, this takes into account the table of a raster dataset is, in fact, an index of the values that uniquely occur in the dataset and the count of cells assigned to each of those values.)*

i. Resampling Method

Raster pixels are square, and the transformations that occur when the software projects raster data or the interpolations the software must account (or not) for portions of a cell or neighboring cells when assigning values to cells in the output dataset. Generically ArcGIS uses a resampling process to determine the best result. As users of the software, we may choose from several resampling methods, each having pros and cons. Esri advises using bilinear resampling when working with spatially continuous data, and nearest neighbor resampling when working with spatially discrete data. A third option of Cubic Convolution is also available. Esri further advises that if the data values in a continuous raster are nominal, i.e., categorical, the nearest neighbor resampling should be used.

In testing the WPE datasets, which are highly discrete, we learned that the level of discreteness matters. We also found that only nearest neighbor preserved the minimum and maximum values occurring in the dataset, meaning bilinear and cubic convolution produced output datasets with new ranges. We were further surprised to find in spite of bilinear resampling slightly changing the range, it produced the least change, just slightly, in total population.

Another way to think of level of discreteness is to evaluate the complexity of the footprint of cells with data values versus the cells set to NoData. Esri developed a method to measure discreteness based on a 25 cells radius where NoData cells nearby a cell carry a higher weight than NoData cells further away. For example, the result was seeing that coastlines, which are inherently complex areas, have lower confidence due to increased exposure to resampling error. Rural areas were similarly affected. Urban areas with large swaths of continuous settlement texture, are relatively continuous, and therefore we have higher confidence resampling does not affect them.Thus, the conundrum is whether to use bilinear resampling, knowing that data values that did not occur in the input dataset, could be interpolated, and potentially change even the range of the output data. The tradeoff is slightly better fidelity with respect to the total population represented. Esri chose to use nearest neighbor resampling with the intent of avoiding the possibility of cumulative errors during the multi-step processes that produce the WPE.

ii. Cell Size

When projecting raster datasets, cell size, in particular, should not be changed from the default setting automatically determined by the Project Raster tool based on the combination of the source and output coordinate systems. We found manually changing the cell size resulted in dramatically exacerbating the impact of resampling, and the more discreet a given raster data set, the worse the effect.

E. Considerations and Range of Applications

The WPE services may be used for a variety of spatial analyses, the most common of which would be to provide an estimate of the number of people living inside a given region or within a distance of a given event such as an earthquake epicenter or along a given segment of coastline. However, qualifying that estimate, in terms of how accurate it might be, is not simple. For instance, the margin of error for such an estimate would need to account for local variation and accuracy of all the input datasets, some of which are not available, and no specified method to do so exists. Such a qualification would also vary on a cell by cell basis making it potentially difficult to express at the level of the polygon representing a region. However, there is information that can be used to gain some idea of the quality of such an estimate, and the following sections provide perspective on using the WPE services.

i. Population Count Values

In the 2015 population estimate layer, the values estimate the count of people living within the area represented by that cell. The values range from 1 to 32,767 in the 162-meter resolution data. Generally, the cells with values of 4 or less are considered very unreliably located. The process of allocating people to sparsely populated areas arbitrarily thins the texture from the Landsat8 Panchromatic imagery (described later in Section III Steps 19-26.) to less than ten percent of the original texture. Depending on where one looks in the world the odds are between one in ten to one in twenty that a single cell four or fewer people is correctly located. The majority of such cells can be said to be in the neighborhood, but may be up to two kilometers, or more from where people may actually live.

As the estimated population for a given cell becomes higher, there is generally reliability. However, at the very top of the range, there are also unreliable cells. This is due to spatial inaccuracies between the polygons reporting, usually urban populations, which can be displaced up to a kilometer. The result is a high number of people allocated to a small amount of footprint. For instance, the upper value of 32,767, represents a population density that physically cannot exist. Esri has been working to progressively eliminate these issues. Thus, consider any value above 5,000 persons in a 162-meter square unreliable unless a very tall building is located within that square. As with the very low values, the total number of people within two kilometers of this location is likely quite accurate.

ii. Confidence Factor

One of the services in the latest version of the WPE is the Confidence level service, which can be found online.

This service represents confidence on the basis of a one to five rating, where one is the least confident and five the most confident. This can be used to either find the average level of confidence for the footprint by taking the mean confidence score for a region, or deriving a function that produces a margin of error for the population estimate of each cell within an area. Note that the larger the region analyzed, the more accurate the estimate is likely to be, as the region may eventually include entire census reporting units, which would represent the highest level of confidence possible. Thus, this discussion is mostly about how much confidence to have in the areas of the world where excellent census data does not exist, and when the population of a relatively small area is being estimated.

iii. Minimum Size of Area to analyze

What is the smallest area that can reliably have its population estimated? Unfortunately, it depends on where that area is located. If the confidence value is high, then the area could be as small as one cell, though it is recommended to use at least one hundred cells in such areas. For areas of least confidence, it is recommended to use a minimum area of 50-100 square kilometers. If confident the values are consistent with their locally informed expectations, then using smaller areas may work.

iv. Density and Anomalies

One way to better understand the probable accuracy of the estimated population values is to plot the cumulative percent of total population by population density (Figure 5).

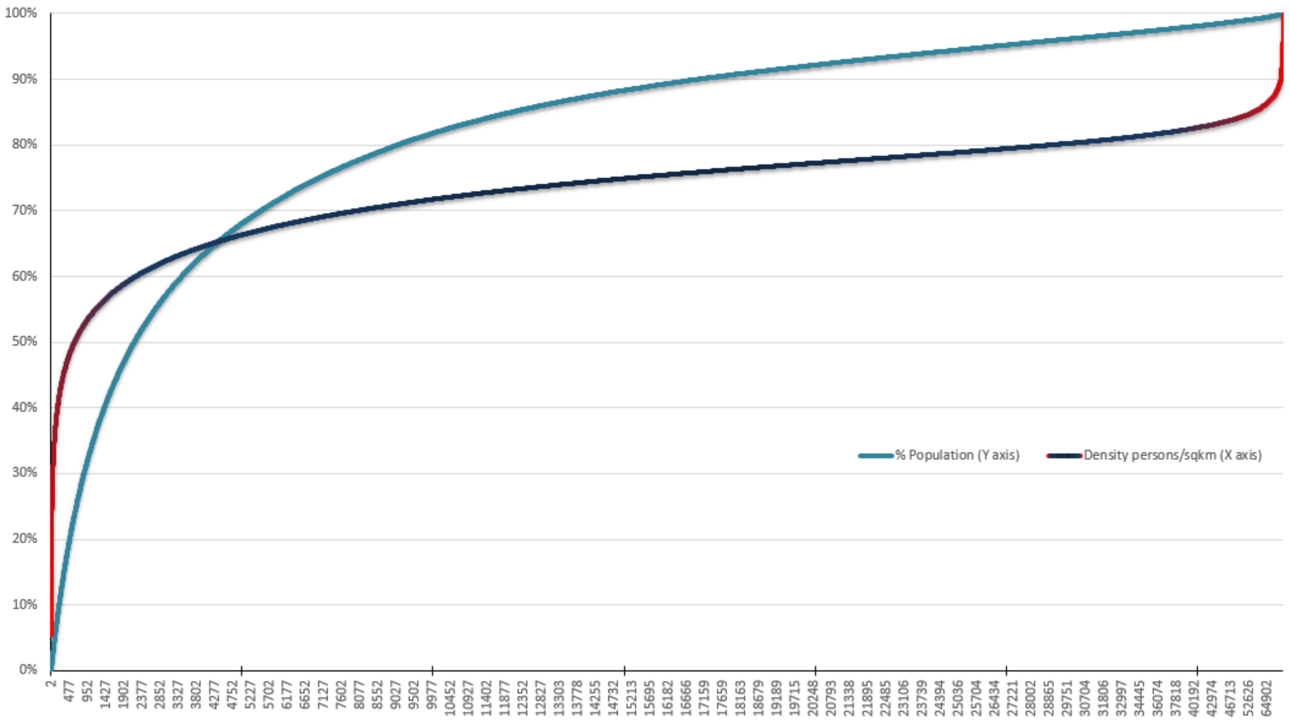


Figure 5. Plot comparing the cumulative percent of total population in blue-green by population density level where density levels red correspond to less reliable values described in section II.E. The purpose is to note the inflection points in the red curve, particularly at the top where the highest densities are likely due to processing error or horizontal accuracy issues in the polygon data used to assign people to the footprint likelihood surface. The inflection point for these upper values can be seen in Figure 4.

III. Methodology

This section will describe the sequence of processing steps, and expectations for each step, used to produce the WPE. The description herein is intended to facilitate replication of the method and collaboration to improve the method.

A. Land Cover Score

The process create the WPE begins with assigning scores for the portions of the landscape that are well known or highly likely. The land cover score is designed as a starting point. Several sources of information are used for this purpose:

- MDA's BaseVue which is a global 30-meter resolution classified land cover product using an Anderson-style classification. BaseVue is also created from Landsat8 imagery. In particular BaseVue as excellent urban footprints at medium-low and high density levels.
- GeoNames place points. There are many small places of settlement that are not dense enough to register in BaseVue, but particularly so for smaller settlements in areas of dense vegetation, or ironically, settlements in areas of very high reflectance, such as deserts where building materials often resemble the surrounding bare earth. These locations help inform the process as to where texture may be valid as opposed to other similar texture nearby.
- Road intersections. Roads lead from settlement to settlement and from homes to settlements. Intersections are a topologically efficient way to enhance a local area's texture score. Road intersections can be efficiently generated in ArcGIS by creating a geometric network from the road line features in a geodatabase—one byproduct of the network is a "Junctions" feature class which can be copied and used separately.

The land cover score is initially calculated as follows:

1. Reclassify the MDA land cover such that the following classes receive the following scores (see Figure 6):
   a. 0: Evergreen Forests, Scrub/Shrub, Barren, Wetland, Mangrove, Snow, Ice, and Clouds.
   b. 25: Agriculture, Grassland (pasture/rangeland), Deciduous forest (orchards)
   c. 150: Medium-low density urban
   d. 200: High density urban
2. Convert the road intersections to a raster at 30m resolution, snapping to the output of step 11. Note the value of each intersection point is preset to 150.
3. Use the result of step 12 as the input to the Block Statistics tool with settings of 5x5 NAW, and Minimum. This spreads the values to 5x5 the neighborhood. One way to think of the intention is that in a rural area, we expect there is an inhabited building somewhere close by.

4. Use the results of Step 13 and Step 11 as the inputs to a Con function that evaluates whether the BaseVue is zero, and if so, override that cell's value to 150. Thus, anywhere near a road intersection will not be omitted in the land cover score.
5. Convert the GeoNames points to a raster at 30m resolution, snapping to the output of step 11. Note the score for each point is 150.
6. Use the result of step 15 as the input to the expand tool and expand by 3 cells. This effectively creates a locale of 180 meter in diameter for each place.
7. Use the results of Step 16 and Step 11 as the inputs to a Con function that evaluates whether the BaseVue is zero, and if so, override that cell's value to 150. The result is the Land Cover score, with values ranging from 0 to 200.

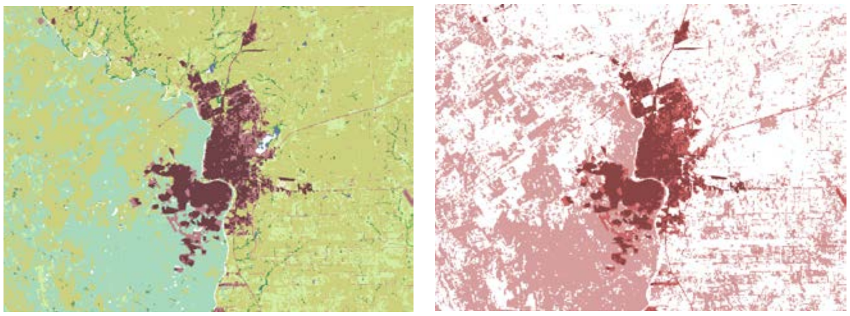


Figure 6. Shows BaseVue 2013 imagery on the left, and the result of Step 1 on the right, where the locations likely to have people living in them are in shades of red, and locations not considered as likely locations for people to live are now white.

B. Landsat8 Panchromatic Texture Score

Next the process adds texture from Landsat 8 Panchromatic images, which are 15-meter resolution raster datasets, called "scenes", distributed by the U.S. Geological Survey (USGS) in partnership with Amazon Web Services (AWS). The purpose of texture is to provide additional variety within the footprint of the land cover score which can then be evaluated and if enough texture is found it may be considered as settlement or if insufficient texture is found, we gain confidence that no people live in that location.

The concept of texture within the Panchromatic Landsat8 imagery was inspired by Haralick's algorithm for identifying textural features within grayscale imagery. In particular, the sum or ranges (shown later) was central. In the course of reviewing this method internally at Esri, staff statisticians were consulted and they noted the similarities between this algorithm and those used in terrain analysis for ecological studies to produce rugosity, or roughness indexes.

We first needed a strategy to process the Landsat8 data. The location of Landsat scenes is consistent, with new scenes added every sixteen days. The locations for scenes are based on a fixed combination of path and row coordinates. More information is available at:

- https://aws.amazon.com/public-data-sets/landsat/
- http://landsat.usgs.gov/l8handbook_section4.php

Esri also publishes a collection of Landsat8 image services, and because that work contained everything needed to process the WPE, and copy of the mosaic dataset, which points to a copy of the Landsat8 imagery maintained by Esri exclusively to support these services, was used. Additionally, Esri obtained scene metadata from the AWS volume to create a score used to choose which of the scenes to show first. This metadata and score were added to the mosaic dataset's attributes, which are at the level of one row per scene. This meta data included the date of the image, the percentage of clouds detected in the image, the path and row, and the sun azimuth and angle. This score, used by the Esri web services, favored scenes with similar dates over levels of clouds. To process the WPE, it was decided that cloud free imagery was important than the dates of adjacent scenes.

The 15-meter resolution Landsat8 panchromatic scenes began accumulating in 2013 and for the 2013 version of the WPE, scenes from the first nine months of available data were used. In order to ensure maximum opportunity for cloud-free scenes, the 2015 version of the WPE used scenes from the thirty-six months. Future versions are expected to use a date range of one year, limited to the previous year of available scenes.

The processing is accomplished using a tile-based workflow with 0.5 degree tiles for the land areas of the world. The tiles are in WGS 1984 and for each tile the data from the Landsat8 scenes with the lowest cloud score was extracted as 16-bit unsigned integer Lempel-Ziv-Welch (LZW) compressed tagged image file format (TIFF) files where values range from 0 (No Data), then skip 1 to 5,499, and then continue from 5,500 to 65535. Because some path and row combinations did not have cloud free imagery, the next steps (below) processed the tile to mask areas with clouds. Note that while the USGS does publish is a quality assurance (QA) image, for each scene available on the AWS site, with a cloud mask, visual analysis showed this was not reliable, in that some clouds were missed, and some highly reflective surfaces were classified as clouds. Therefore, the QA imagery was not used. Note that the CF imagery from Landsat8 is not on the AWS site, and still needs to be tested as a potential substitute for the next steps.

i. Cloud Processing

8. All values in the tile above 15,000 are initially considered cloud and extracted via a Con statement Con("%LS8Tile%" > 15000,1,0) to new separate dataset. Note values of 1 represent clouds and zero non-cloud.
9. These cells are processed using the Region Group tool set to four neighbors using the Within Zone Group Method. This produces a unique ID for each contiguous group of cells.
10. Resulting Cell values with extremely low or extremely high cell counts relative to the number of cells per tile are excluded via a Con statement: Con ((Count >= 100 AND Count <= 100000), "%LS8Tile%"). Tiny clouds or more likely small glints from water bodies and buildings are eliminated, as well as large areas such as salt flats, playas, irrigated agriculture, etc. are now no longer considered to be clouds.
11. The resulting footprint of clouds is expanded by twenty cells using the Expand tool.

12. The original extracted Landsat8 tile dataset is processed to set the cells with clouds (result of step 4) to NoData using the SetNull tool.

There are pros and cons to this approach.

First the cons:

- Human settlement, particularly cities, contain values over 15,000 due to the relatively high number of reflective surfaces in the built environment.
- The expand function, which is intended to mask cloud shadows, dramatically excludes data, creating large holes in cities while at the same time removing areas showing clouds.
- The region group tool is by far the slowest part of the process.

The pros:

- A great deal of what is eliminate is unpopulated. However, when areas of city are eliminated, the land cover score (detailed later) overrides in these areas, so the prospect for producing false negative results is mitigated.
- Visual inspection of the panchromatic images shows the QA imagery does not reliably classify all clouds. Steps 1-5 above are more reliable.

ii. Initial Texture Score Processing

The texture score indicates how much local variety, within a 5x5 cell neighborhood (about one city block) occurs. The premise is the higher the variety, the more likely human settlement at this location.

13. The result of step 5 is the input to the Focal Statistics tool which calculates the Range (max – min) of cell values within a 5x5 cell moving neighborhood area window (NAW). This indicates how much local variance of cells values exists. See Figures 7 and 8.
14. The result of step six is the input to the Focal Statistics tool which calculates the Sum of cell values within a 5x5 cell moving NAW. The high values indicate a high local variety that it is not due to one-cell anomalies. See Figures 7 and 8.
15. Filter to isolate the initial footprint based only on texture scoring. If the values in the result of step 7 are greater than the half degree tile's mean (μ) plus one standard deviation (σ), then the cell is considered part of the footprint.
16. Generalize the results. Because steps 6-8 are best at identifying the edges of change, we need to insure the areas within one city block are included with the edge. To do that the result of step 8 is the input to the Focal Statistics tool, which calculates the mean (μ) of a 4x4 cell moving NAW to produce the initial texture score.
17. Normalize the scores to values range from 0 (uninhabited) to 100 (most likely to be inhabited), by multiplying the log of the sum of ranges above threshold by 7.0.

The initial texture score can be expressed as follows:

$$T_c = \sum_{i=c}^{25} (C_{Max} - C_{Min})_{\, n=5x5} \qquad T_{Settlement} = T_c > (T\mu_s + T\sigma_s)$$

Where: T = Texture, c = cell location, n = neighborhood area window, s = 0.5 degree processing tile

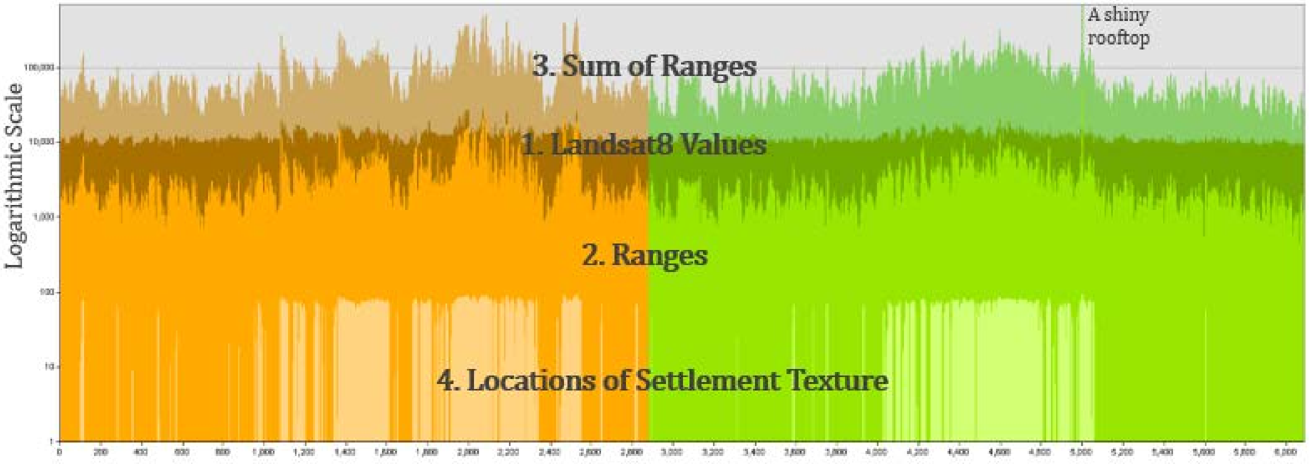


Figure 7. A graph of profile paths shown in Figure 8. This illustrates which cells are selected as settlement, which can be verified by looking along the profile lines (from left to right).

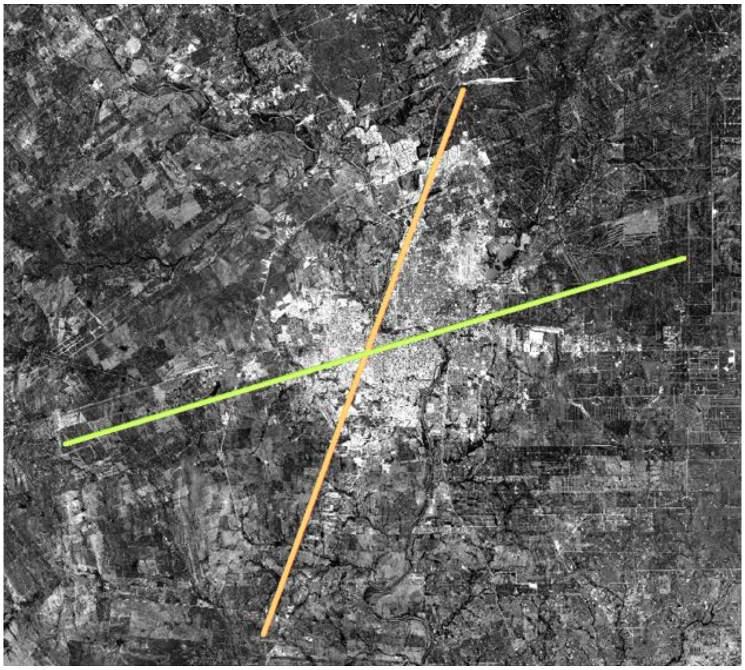


Figure 8. Profile lines matching the color scheme of Figure 7. The area shown is of Laredo, Texas in the U.S. to the northeast, and of Nuevo Laredo in Mexico to the Southwest.

Steps 6-9 find texture we most associate with human settlement, but also find other circumstances with the same statistical profile. Some settlement texture is missed due to the coarse cell size of 15-meters. Figure 9 depicts several prototypical texture scenarios representing texture values in adjacent raster cells in the panchromatic Landsat8 imagery.

One issue that will be tested for the 2016 process is the mosaic dataset that references the panchromatic Landsat8 scenes will have a correction for top of atmosphere radiance. The theory is doing so will better normalize the values on a per scene basis, allowing global thresholds (versus the tile's statistics).

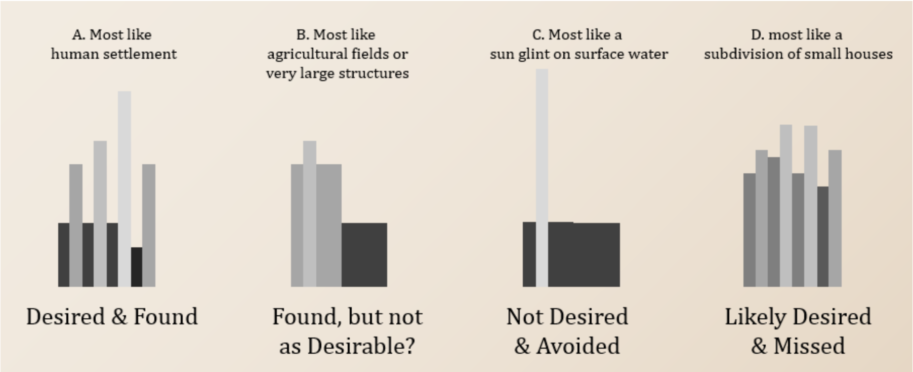


Figure 9. These hypothetical profile sequences for Landsat8 panchromatic cell values illustrate: A) desirable texture indicating human settlement based on the lighter values indicating tops of buildings and darker areas the shadows cast by those buildings, B) texture with the same summary statistics, C) texture that is avoided by using the sum or ranges, rather than only the range of values within the 5x5 NAW in step 6, D) texture that is missed because the local range is too low.

C. Final Footprint Score

The final footprint score is accomplished in two stages. The first combines the Landsat8 texture score to the land cover score as follows:

18. Using the results of Step 10 (Landsat8 texture score) and Step 17 (Land cover score) as inputs to a Con function such that if the land cover score is greater than 0, then the output should be the sum of the land cover and Landsat8 texture scores:
    Con("%LCTile%" > 0,("%LCTile%" +"%LS8Tile%"))

The second stage addresses the problem that can occur when apportioning population from a polygon reporting unit. The problem is when the cell with the highest score cannot be allocated with at least one person. This occurs typically occurs when there are more cells in the footprint than people to allocate. Population must be allocated in integers, which cannot be split across cells.

19. Convert the population polygons from the census data (fully described in the next section) to raster such that the value for the resulting raster dataset is the polygon's object ID, which will serve as unique identifier (to ensure that different polygons with the same population are not used in a conjoined fashion).

20. Use the results of step 18 in a Con function where if the value is greater than zero, it is set to one, otherwise Null.
21. Use the result of step 19 and the result step 20 as inputs to the Zonal Statistics to Table tool such that the results of Step 19 are the zones and the statistic will be a sum and a count of potentially populated cells within each population polygon, because the values are now one.
22. Use the Join field tool to add the count of cells field and the sum of scores field to the population polygons.
23. Use an Attribute query on the Population polygons to select polygons where the count of cells divided by the maximum score from the result of Step 18 is less than 1.0. This selects all polygons where fewer than a whole integer person could be allocated to a single cell.
24. Keeping the selection from step 23, convert the population polygons to a raster such that the population is the value.
25. Use the results of step 24 in a Con function where if the value is greater than zero, it is set to one, otherwise Null. This will be a processing mask for the next step.
26. Use the results of step 18 and 25 as the inputs to a Con function such that if the result of step 25 is Null, then use the result of step 18, Else if the score from Step 18 is less than 124 then set it to zero, or if the score is greater than 123, subtract 124 and multiply that by 1.73.

D. Allocation of Population

27. Using the result of Step 22, clear the selection, and use as the input to Polygon to Raster, create a raster based on the sum of scores field.
28. Use the population polygons as input to the Polygon to Raster tool and create a raster with the value coming from the population field.
29. Using the result of Steps 26, 27, and 28 in a Raster Calculator with this function: Int(((Float("%Step26%") / Float("%Step27%")) * Float("%Step28%"))). This translates to, for each cell divide the score by the sum of scores for the polygon (to give a percent of total for that cell within the polygon), and multiply that by the population for the polygon.

Each of the 0.5-degree tile rasters were added to a Mosaic dataset that was then converted to one global raster dataset to be used in production and services.

The above steps are predicated on a thus far undescribed population polygon layer. This layer is an amalgamation of many sources, with the intent of using the most current publicly available or available at low cost census or census surrogate data applied to the finest levels of geography possible. One of the goals of each new version of the WPE is to improve upon this by adding more polygons representing progressively finer tabulation geographies. The current population polygons dataset contains data from the following sources:

- United States: U.S. Census Block Group with Esri's current year estimate
- Canada: Environics Analytics at the Dissemination Area (DA) level.
- Michael Bauer Research GmbH: 130 countries at Admin level 3 (county) or 4 (city/town) for current year estimate
- United Nations most recent estimate in all other cases usually at admin level 3 though some are 2 (state).

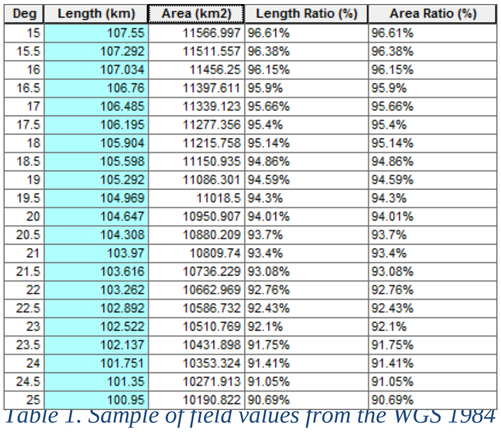
One issue affecting this step is the horizontal accuracy of the population polygons relative to the resolution of the raster data and location of cells with settlement texture.

E. Aggregate to 150-meter resolution

We chose to aggregate the result of Step 29 data to ~162-meter resolution to facilitate analysis from a web service, which is ultimately subject to bandwidth constraints. ArcGIS image services have a parameter to restrict the size of an image a client may request in order to protect the server's resources from being over used by a single client, and thus curtailing use by other simultaneous clients. This is also based on how long it takes the server to process an image of a given size. Based on the server's hardware, we set a maximum normal processing time limit of ten seconds, which is smaller than the server's timeout setting of thirty seconds, which effectively allows two concurrent users per instance of a server to make such a request. At a resolution of 15 meters, an area of only several hundred square kilometers could be allowed. Shifting the output resolution to ~162 meters permitted an area the size of Africa to be analyzed directly from the service. The Aggregate tool with a cell factor of 10 set to Sum the values was used as a final step.

F. Deriving Population Density

Population Density was derived by multiplying by the cells in another raster dataset (Table 1 and Figure 10) that represented the percent of area a given cell has relative to a cell at the equator. Thus cells at the Equator are 100%, and cells at the extreme latitudes are single digit



Deg	Length (km)	Area (km2)	Length Ratio (%)	Area Ratio (%)
15	107.55	11566.997	96.61%	96.61%
15.5	107.292	11511.557	96.38%	96.38%
16	107.034	11456.25	96.15%	96.15%
16.5	106.76	11397.611	95.9%	95.9%
17	106.485	11339.123	95.66%	95.66%
17.5	106.195	11277.356	95.4%	95.4%
18	105.904	11215.758	95.14%	95.14%
18.5	105.598	11150.935	94.86%	94.86%
19	105.292	11086.301	94.59%	94.59%
19.5	104.969	11018.5	94.3%	94.3%
20	104.647	10950.907	94.01%	94.01%
20.5	104.308	10880.209	93.7%	93.7%
21	103.97	10809.74	93.4%	93.4%
21.5	103.616	10736.229	93.08%	93.08%
22	103.262	10662.969	92.76%	92.76%
22.5	102.892	10586.732	92.43%	92.43%
23	102.522	10510.769	92.1%	92.1%
23.5	102.137	10431.898	91.75%	91.75%
24	101.751	10353.324	91.41%	91.41%
24.5	101.35	10271.913	91.05%	91.05%
25	100.95	10190.822	90.69%	90.69%

Table 1. Sample of field values from the WGS 1984 Density Conversion raster (Figure 9), showing conversion factors for latitude range 15 to 25 deg. in either direction from the equator.

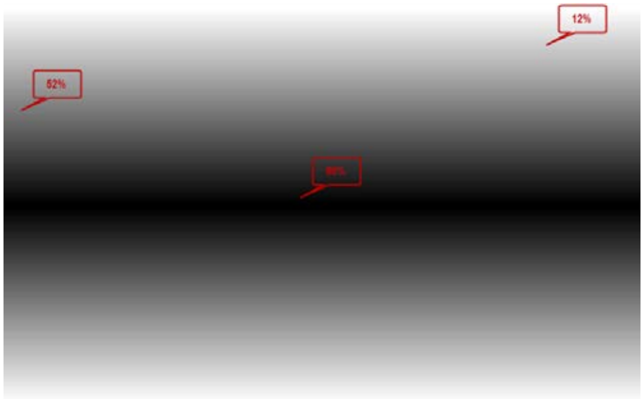


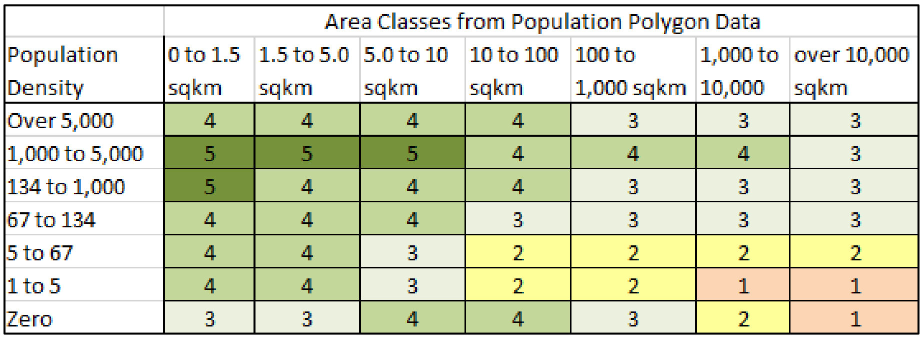
Figure 10. WGS 1984 depiction of the percentage of area a cell represents the farther away from the equator it is located.

percentages. The raster dataset with the percentages has a resolution of 0.5 degrees, and was considerably faster than a cell-wise processing of the population raster, which could not be completed on any computer we had (Doing so requires the Zonal Tabulate Area tool, which could not complete the job on a global raster of even 150-meter resolution, much less at 15-meter resolution. The 0.5-degree dataset is available online:

G. Confidence Surface

There is currently a confidence surface based on the average of two factors. The first is the ratio of the area of the population polygon to the number of people (Table 2), and the second is a score for how complex the footprint is relative to NoData and zero population cells.

i. Population to Area Confidence Scoring



	Area Classes from Population Polygon Data						
Population Density	0 to 1.5 sqkm	1.5 to 5.0 sqkm	5.0 to 10 sqkm	10 to 100 sqkm	100 to 1,000 sqkm	1,000 to 10,000	over 10,000 sqkm
Over 5,000	4	4	4	4	3	3	3
1,000 to 5,000	5	5	5	4	4	4	3
134 to 1,000	5	4	4	4	3	3	3
67 to 134	4	4	4	3	3	3	3
5 to 67	4	4	3	2	2	2	2
1 to 5	4	4	3	2	2	1	1
Zero	3	3	4	4	3	2	1

Table 2. Shows a matrix of confidence scores based on the population of a given reporting polygon and its area in square kilometers. The range is 5 for highest confidence and 1 for lowest confidence.

ii. Footprint Complexity Score

The footprint complexity was computed as a per cell score that rated the number and distance to no data cells within 8 kilometers. The reason complexity is a problem is due to necessary raster resampling processes that occur during the WPE projection workflow. Edges of data and NoData cause allocation errors between raster datasets and between raster and vector datasets, typically resulting in underestimation. The processing steps are as follows:

1. Project the WGS_1984 output of the 150-meter resolution aggregation to World Equidistant Cylindrical. This ensured the distances between cell centers would be comparable.
2. To speed the processing, the result of Step 1 was aggregated by a factor of 2 cells with a sum operation.
3. Reclassify the result of step 2 where values of zero or NoData are set to NoData, and all other values are set to 1.
4. Use the result of Step 3 as the input to the Euclidean Distance tool. The divide that result by 1000, to convert it to kilometers.

5. Use the Get Raster Properties to learn the Maximum value from the result of Step 4.
6. Use the result of Step 3 as the input to Focal Statistics, set to a Circular Neighborhood with a radius of 8 cells, and a Sum operation.
7. Divide the Result of Step 6 by the result of Step 4. This produces a ratio of the count of data cells within 8 kilometers to the distance to the nearest data cell
8. Reclassify the result of step 7 based on a 5-Class Natural Breaks distribution such that the low values receive a value of 1 and the high values a value of 5.

We used the confidence scores as zones for the Zonal Statistics to table tool and summed the population within each confidence level (Table 3).

Confidence Level	Count of people	% of Total population	% Urban Density
1 = Least	32,679,164	0.45%	0.85%
2	1,406,643,023	19.30%	20.12%
3	2,506,831,960	34.40%	35.74%
4	2,720,870,831	37.33%	36.96%
5 = Most	620,957,914	8.52%	6.33%

Table 3. Shows the counts, percent of total, and urban percentages at each confidence level.

Additional factors are planned for future editions of the confidence surface, including a score for the quality of the census information that provides the population figure to each populated polygon. The idea is for the best de-jure censuses to have a confidence of 5, while the countries without censuses would have a value of 1, all others may have a value of three unless the United Nations Statistical Division has a quality score to use. A second metric to include would be to factor in local variance of scores, which should be relatively low.

IV. References

The following references were used in the course of deriving and confirming the method presented herein for producing the WPE.

Azar, D Graesser, J Engstrom, R, Comenetz, J Leddy, R M Schechtman, N G and **Andrews T** 2010 Spatial Refinement of Census Population Distribution Using Remotely Sensed Estimates of Impervious Surfaces in Haiti" *International Journal of Remote Sensing* 31.21 (2010): 5635-655. Web.

Balk, D 2003 "Improving Global Population Estimates" In: Conference on Migration, Urbanization, and Health, Princeton University, September 25, 2003.

Bhaduri, B and **Bright, E** 2003 LandScan Population Projects. In: Conference on Migration, Urbanization, and Health, Princeton University, September 25, 2003.

Bhaduri, B Bright, E Coleman, P and **Urban, M L** 2007 LandScan USA: A High-resolution Geospatial and Temporal Modeling Approach for Population Distribution and Dynamics. *GeoJournal* 69.1-2 (2007): 103-17. Web.

Bielecka, E 2005 A Dasymetric Population Density Map of Poland. In: International Cartographic Conference, A Coruña, Spain, August 2005.

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 (2007): 119-29. Web.

Encarnação, S, Gaudiano, M Santos, F C Tenedório, J A and **Pacheco, J M** 2012 Fractal Cartography of Urban Areas. *Scientific Reports* 2 (2012): n. Web.

Encarnação, S, Gaudiano, M Santos, F C Tenedório, J A and **Pacheco, J M** 2012, Fractal Cartography of Urban Areas: Supplementary Information. *Scientific Reports* 2 (2012): Web.

Esch, T Taubenbock, H Roth, A Heldens, W Felbier, A Theil, M Schmidt, M Muller, A A and **Dech, S** 2012 TandDEM-X Mssion- New Perspectives for the Inventory and Monitoring of Global Settlement Patterns. *Journal of Applied Remote Sensing* 6 (2012): 0617021-06170221. Web.

Esch, T Marconcini, M Felbier, A Roth, A Heldens, W Huber, M Schwinger, M Taubenbock, H Muller, A A 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 (2013): 1617-621. Web.

Esch, T Marconcini, M Marmanis, D Zeidler, J Elsayed, S Muller, A A and **Dech, S** 2014 Dimensioning Urbanization- An advanced procedure for characterizing human settlement properties and patterns using spatial network analysis. *Applied Geography* 55 (2014): 212-228. Web.

Getchee Inc. 2011 *A Quick View of Grid Demographic Processing. Web*

Haralick, R M, **Shanmugam, K** and **Dinstein, I** 1973 *Textural Features for Image Classification* IEEE Transactions on Systems, Man and Cybernetics Vol. SMC-3 No. 6 Nov 1973 pp 610-621.

Kim, B Youkyung, H Yonghyun, K and **Yongil, K** 2014 Generation of Cloud-free Imagery Using Landsat-8. *Journal of the Korean Society of Surveying Geodesy Photogrammetry and Cartography* 32(2):133-142, April 2014

Linard, C and **Tatem, A J** 2012 Large-scale Spatial Population Databases in Infectious Disease Research. *International Journal of Health Geographics* 11.1 (2012): 7. Web.

Linard, C Gilbert, M and **Tatem, A J** 2012 Assessing the Use of Global Land Cover Data for Guiding Large Area Population Distribution Modelling. *GeoJournal* 76.5 (2011): 525-38. Web.

Maantay, J Maroko, A and **Herrmann, C** 2007, Mapping Population Distribution in the Urban Environment: The Cadastral-based Expert Dasymetric System (CEDS). *Cartographic and Geographic Information Science* 34.2 (2007): 77-102. Web.

Miyamota, E and **Merryman T Jr** 2005 Fast Calculation of Haralick Texture Features Unpublished Classroom Materials accessed from https://users.ece.cmu.edu/~pueschel/teaching/18-799B-CMU-spring05/material/eizan-tad.pdf

Patterson, L Urban, M Myers, A Bhaduri, B Bright, E A and **Coleman, P R** 2009 The Effects of Quality Control on Decreasing Error Propagation in the LandScan USA Population Distribution Model: A Case Study of Philadelphia County. *Transactions in GIS* 13.2 (2009): 215-28. Web.

Pesaresi, M, Blaes, X Ehrlich, D Ferri, S Gueguen, L Haag, F Halkia, M Heinzel, J Kauffmann, M Kemper, Ouzounis, T O 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, 2012. Web. http://publications.jrc.ec.europa.eu/repository/handle/JRC77925.

Sabesan, A Abercrombie, K Ganguly, A R Budhendra, B Bright, E A and **Coleman, P R** 2007 Metrics for the Comparative Analysis of Geospatial Datasets with Applications to High-resolution Grid-based Population Data. *GeoJournal* 69.1-2 (2007): 81-91. Web.

Sleeter, R 2014 Dasymetric Mapping Techniques for the San Francisco Bay Region, California In: URISA 2014.

Tang, J 2003 Evaluating the Relationship between Urban Road Pattern and Population Using Fractal Geometry. UCGIS.org Web: http://www.ucgis.org/summer03/studentpapers/junmeitang.pdf

Tatem, A J Adamo, A Bharti, N Burgert, C R Castro, M Dorelien, A Fink, G Linard, C Mendelsohn, J Montana, L Montgomery, M R Nelson, A Noor, A M Pindolia, D Yetman, G and **Balk, D** 2012 Mapping Populations at Risk: Improving Spatial Demographic Data for Infectious Disease Modeling and Metric Derivation. *Population Health Metrics*10.1 (2012): 8. Web.

Taubenbock, T Esch, T Felbier, A Wiesner, M Roth, A and **Dech, S** 2012 Monitoring urbanization in mega cities from space. *Remote Sensing of Environment* 117 (2012): 162-176. Web.

Taubenbock, T Esch, T Roth, A and **Dech, S**, 2011 Pattern-Based Accuracy Assessment of an Urban Footprint Classification Using TerraSAR-X Data. *IEEE Geoscience and Remote Sensing Letters* 8.2 (2011): 278-282. Web.

Zhou, Y Yang, G Wang, S Wang, L Wang F and **Liu, X** 2014, A new index for mapping built-up and bare land areas from Landsat-8 OLI data. *Remote Sensing Letters*, 5:10, 862-871, DOI: 10.1080/2150704X.2014.973996

I. Documentation Copyright and License

Copyright 2017. Esri

Appendix. Data Revision History

2013 Edition: Initial creation
2015 Edition: Updated Population estimate polygon datasets, road intersections, and Landsat8 panchromatic imagery.