Multitemporal settlement and population mapping from Landsat using Google Earth Engine

Nirav N. Patel[a,*], Emanuele Angiuli[b] , Paolo Gamba[b], Andrea Gaughan[c], Gianni Lisini[b], Forrest R. Stevens[d], Andrew J Tatem[e,f,g] , and Giovanna Trianni[b]

a. Department of Geography and Geoinformation Science, George Mason University, 4400 University Drive, MS 6C3, Fairfax VA 22030 USA

b. Department of Electrical, Biomedical and Computer Engineering, University of Pavia, Italy

c. Department of Geography and Geosciences, University of Louisville, 213 Lutz Hall, Louisville, Kentucky 40205 USA

d. Department of Geography, Land Use and Environmental Change Institute, University of Florida, Gainesville, Florida 32609 USA

e. Department of Geography and Environment, University of Southampton, Highfield, Southampton SO17 1BJ, UK.

f. Fogarty International Center, National Institutes of Health, Bethesda, MD 20892, USA.

g. Flowminder Foundation, 17177 Stockholm, Sweden

*Corresponding author at: 13706 Sun Court, Tampa, Florida, 33624, Tel: +01 813-766-2801

Email address: niravpatel65@gmail.com (N.N. Patel)

Abstract

As countries become increasingly urbanized, understanding how urban areas are changing within the landscape becomes increasingly important. Urbanized areas are the often

the strongest indicators of human interaction with the environment, and understanding how urban areas develop through remotely sensed data allows for more sustainable practices. The Google Earth Engine (GEE) leverages cloud computing services to provide analysis capabilities on over 40 years of Landsat data. As a remote sensing platform, its ability to analyze global data rapidly lends itself to being an invaluable tool for studying the growth of urban areas. Here we present (i) an approach for the automated extraction of urban areas from Landsat imagery using GEE, validated using higher resolution images, (ii) a novel method of validation of the extracted urban extents using changes in the statistical performance of a high resolution population mapping method. Temporally distinct urban extractions were classified from the GEE catalog of Landsat 5 and 7 data over the Indonesian island of Java by using a Normalized Difference Spectral Vector (NDSV) method. Statistical evaluation of all of the tests were performed, and the value of population mapping methods in validating these urban extents were also examined. Results showed that the automated classification from GEE produced accurate urban extent maps, and that the integration of GEE derived urban extents also improved the quality of the population mapping outputs.

Keywords: Landsat, Multitemporal, Population Mapping, Google Earth Engine, Settlement Mapping, Urbanization, spatial demography

1. Introduction

Landsat imagery have proven to be useful in understanding global urbanization trends over different timescales. Satellite-derived data have been integral in understanding trends in urban sprawl and many other dynamics of urbanization (Guindon et al., 2004; Angel et al., 2005; Burchfield et al., 2006; Schneider & Woodcock, 2008; Potere et al., 2009; Schneider, 2012; Taubenböck et al., 2012; Sexton et al., 2013).

The Google Earth Engine (GEE) is an online environmental data monitoring platform that incorporates data from the National Aeronautics and Space Administration (NASA) as well as the Landsat Program. After the USGS opened access to its records of Landsat imagery in 2008, Google saw an opportunity to use its cloud computing resources to allow records of Landsat imagery to be accessed and processed over its online system. This has enabled users to reduce processing times in analyses of Landsat imagery and make global scale Landsat projects more feasible (e.g. Hansen et al., 2013). The 30m spatial and multi-spectral resolution is ideal for defining urban areas, and its revisit time is sufficient for monitoring applications (Woodcock et al., 2008). Moreover, because of Landsat's temporal continuity from 1972 to the present day, it is a popular platform to use for urban change analysis (Alberti et al., 2004; Bagan & Yamagata, 2012; Rawashdeh & Saleh, 2006; Yuan et al., 2005).

In the past two decades, the Landsat platform has been paired with imagery from the Advanced Very High Resolution Radiometer (AVHRR) (Hansen et al., 1998), the Defense Meteorological Satellite Program's Operational Linescan System's nighttime imagery (Elvidge et al., 1996, 1997, 1999; Sutton, 2003), and NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) (Schneider et al., 2003, 2009, 2010) to improve the accuracy of urban detection and mapping across large areas. The improvement of methods for detecting urban extents has also driven improvements in population mapping. Satellite imagery has formed the basis of many large area population mapping efforts, such as the Global Rural–Urban Mapping Project (Center for International Earth Science Information Network (CIESIN), 2004) LandScan (Bhaduri et al., 2007) and WorldPop (Linard et al. 2012, Gaughan et al. 2013, "The WorldPop Project," 2014). Satellite-derived urban extents and, more generally, land cover tend to form an important component of accurate population mapping (Linard & Tatem 2012, Linard

et al., 2011), but detailed data can often be costly or time consuming to produce. The GEE presents the possibility for analyzing and classifying satellite data with great speed, so that more relevant and accurate outputs in terms of distributions of population can become a reality (Hansen et al., 2013).

Here we present an approach for the automated extraction of urban areas from Landsat imagery built into the GEE, and a novel method of validation of this mapping using changes in the statistical performance of a high resolution population mapping method (Stevens et al., 2014).

2. Methods

2.1 Study Area

The study area is the Indonesian island of Java, which, along with being the world's most populous island, is also only the fourth largest island in Indonesia but contains more than half of the island nation's population. Jakarta, the capital city is also located on the island and is Indonesia's largest city. The island is 661 miles long from east to west, it ranges in width from about 60 miles in the center to more than 100 miles near each end (Fig. 1).

2.2 Urban Extent Extraction Procedure

The urban extraction methodology proposed here is based on supervised classification of multispectral data. In this work we consider "urban areas" all the portion of a scene with spectrum similar to selected training areas. These training areas include buildings, roads and other artificial surfaces. Therefore, in the following "urban extents" do not correspond to "built-up extents". Our definition of urban areas is instead more similar to "impervious surfaces".

Accordingly, the implemented processing chain is a spectral-based analysis followed by a spatial regularization that is undertaken using the Google Earth Engine cloud computing environment. Processing and implementation in a cloud environment allows for a consistent scaling of the computational efforts when dealing with wide geographical areas. The extraction procedure includes three steps, briefly detailed below: (i) preprocessing and selection of a set of Landsat scenes covering the geographical area and time span of interest, (ii) computation of the Normalized Difference Spectral Vector index (NDSV), a collection of spectral indices that have already been proven (Angiuli & Trianni, 2014) to be an efficient input to urban extent classification algorithms classification and, (iii) spatial-based post-processing.

2.2.1 Preprocessing and scene selection

Pre-processing include orthorectification and coregistration of all the scenes, so that data acquired at multiple dates overlap. This is done internally and seamlessly by the GEE platform at the ingestion of the data from the USGS repository. No radiometric intercalibration or atmospheric correction is performed however. Therefore, although all scenes are calibrated according to the sensor parameters, some differences in radiance values due to the illumination and atmospheric conditions still affect overlapping regions among scenes.

Scene selection is instead performed by our algorithm. Specifically, in order to reduce the Landsat data set to the most suitable scenes, a filter on scene parameters is first applied, to consider only those with less than 10% of cloud coverage and the highest radiometric quality.

2.2.2 Implementation of the Normalized Difference Spectral Vector Index Stack into the Google Earth Engine

Unlike threshold-based recognition of human settlement (one index) approaches developed by Pesaresi et al. (2008) and Xu, (2008), the main input to the urban extent extraction outlined here is the Normalized Difference Spectral Vector (NDSV), proposed in the technical literature (Angiuli & Trianni, 2014) as a means to group existing normalized difference indices (such as the Normalized Difference Vegetation Index - NDVI, the Normalized Difference Water Index - NDWI, and the Normalized Difference Built-up Index - NDBI). NDSV includes in one single vector all the possible normalized indexes that can be computed starting from a Landsat 5 or 7 image, considering therefore 6 bands and 15 possible combinations (the dual ones are not considered as their result is the same but with just the opposite sign).

NDSV includes in one single vector all the possible normalized indexes that can be computed starting from the 30 m spatial resolution bands a Landsat 5 or 7 image. For each band pair this is computed:

$$(1)\ NDSV_{ij} = \frac{b_i - b_j}{b_i + b_j}$$

Hence, using 6 bands and applying Eq. (1) to any possible pair of different bands, a total of 30 indexes are obtained. Due to the symmetry of the definition, 15 of them are only the negative of the other ones, and can be discarded. Each pixel is thus characterized by a set of values, some of which correspond to known indexes (e.g., $NDSV_{43} = NDVI$, $NDSV_{42} = NDWI$, $NDSV_{45} = NDBI$), while other ones have not been explored so far.

Each pixel is thus characterized by a set of values that have been at this point “labeled” only partially. Considering a radiometrically and geometrically corrected Landsat scene, the NDSV features characterizing urban areas, compared to other classes, are shown for a few sample pixels in Fig. 2. It can be noted that urban areas exhibit a distinct NDSV spectral

signature which can be discriminated from other classes by their distinct behavior in this new "multispectral" 15-dimensional space. Fig. 2 demonstrates NDSV profiles that can be obtained from an image.

In summary, instead of relying on threshold-based recognition of human settlements according to a single index ((Pesaresi et al., 2008) and (Xu, 2008)), the procedure implemented in this work considers more information as input to a suitable classification chain, aimed at providing a consistent methodology that works in many different environments, and is reasonably robust with respect to the date of acquisition of the image and unaffected by differences in spatial patterns.

2.2.3 Processing of multitemporal urban extents over Java

Four tests were conducted in order to validate the creation of urban extents using the procedure discussed in the preceding subsections. A census-based population disaggregation method was used for validation, a method that rasterizes GIS data and distributes population counts based on the GIS data that is provided. This method was used because it provides the ability to analyze how the urban extents improve the statistical correlations in the disaggregation process.

In three of four tests, the urban extents were considered as one of the inputs to a census-based population disaggregation method (Stevens et al., 2014). In the first test, instead, the same method was run using the original data sets detailed in Table 1 and the landcover map, including urban extents, was taken from the EarthSat Geocover land cover thematic mapper-based dataset (2007, 30m) by MDA Federal (MDA Federal Inc., 2007). Test 1 served as the baseline data for

validation because it does not use any Google Earth Engine urban extents and because its classification has been validated by MDA Federal, serving as a useful control test.

The qualitative differences of these different landcover-based experiments:

- Test 1: EarthSat GeoCover Landsat Thematic Mapper (TM) derived land cover data from MDA Federal (2007)
- Test 2: GEE urban extents for Java derived using three collections: Imagery from 2006, 2007 and 2008 merged with GeoCover
- Test 3: GEE urban extents for Java derived using three collections: Imagery from 2009 T1 (January through April), 2009 T2 (May through August), 2009 T3 (September through December) merged with GeoCover
- Test 4: GEE urban extents for Java derived using three collections: Imagery from 2008, 2009, 2010 merged with GeoCover

The GEE urban extractions were obtained using Landsat 5 or Landsat 7 data sets, because both satellites were operative in the years of interest. Specifically, multiple Landsat images in the same area and covering a finite period of time were combined in a so called GEE collection, and each pixel was assigned the median value for all images where it appears. Collections are a powerful way to get rid of many of the cloud contaminated pixels, because clouds do not appear in the same position in all images. A better approach would be to mask cloud pixels with a dedicated filter, a function which unavailable in GEE. Therefore, although we understand that cloud-contaminated pixels may still be present in areas with consistent cloud coverage along the year, this technique was assumed as the best available option. Additionally, it must be noted that collections change the radiometric properties of the data, reducing the effectiveness of the

proposed urban extent procedure. To reduce this effect, urban extents for one year were obtained by subdividing the year into thirds. Computing collections for each of these time periods involved extracting urban extents and then combining the resulting maps by majority voting. Similarly, three year collections were subdivided into thirds (one for each year) and then combined by majority voting. To prove the usefulness of the proposed approach for mapping urban extents (and derive population counts) along multiple years, the fourth test repeats the approach of the third one, but using Landsat data collected two years later (2009 versus 2007).

2.2.4 Post-processing

Human settlements can be characterized by peculiar spatial patterns, however, it is important to include a post-processing step aimed at reducing issues related to misclassifications at the pixel level. The simplest and most effective approach is to include morphological operators aimed at discarding isolated pixels and at improving the homogeneity of the extracted settlements with respect to their spatial distribution. Additionally, as the classification results may be affected by spectral patterns (and sub-pixel mixing problems) similar to urban ones in water bodies with high turbidity (Carpenter & Carpenter, 1983; Foody, 2000), such as inner reservoirs, coastal areas and river estuaries, these zones are automatically masked out from the classification in GEE using ancillary GIS data. Similar issues may be caused by clouds, and thus "cloud removal" approaches had to be considered.

2.3 High Resolution Population Mapping Method

As mentioned above, the population mapping algorithm in Stevens et al. (2014) is an essential portion of this study. Thus, its processing steps are briefly described in the following paragraphs.

2.3.1 Population data grid

The 130 census polygons for Java (Figure 1) contained population counts from the year 2010. The population mapping algorithm outlined in Stevens et al (2014) was used, where census counts from the census year are redistributed according to weights, then adjusted up/down based on rural and urban growth rates to a particular year of interest (2007 in this case). This is usually based on the classified urban/rural land cover (built pixels are classified as urban vs. rural using Schneider, et al. (2010) urban/rural MODIS-derived classifications), but in this circumstance uses the new GEE-derived urban delineations to identify urban built pixels. The urban/non-urban delineation was integrated into the MDA landcover data as "built" areas ("_BLT"). The particular year of interest that was selected was 2007 for all datasets, to pick one year for counts to match and for a point of comparison for the accuracy assessment detailed in 2.3.3.

The administrative units were used to delineate the areas where the landcover data in continuous raster format and converted vector format are interpolated by means of the Random Forest method to generate a weighting layer (Stevens et al 2014). Once this weighting layer is generated, population counts for each census unit are distributed over the weighting layer to provide a map of population counts at a 100 by 100 meter resolution (See Table 1 for detail on all covariate datasets used in the process).

2.3.2 Data preparation and the Random Forest population disaggregation method

The general process used for the data preparation, modeling and validation for the population mapping is outlined in Fig. 3. Full details on these steps are provided in Stevens et al., (2014). In brief, the steps in green represent the data preparation tasks. The aggregated population counts and the raster and vector layers shown in Table 1 are then used to create a

Random Forest model (Breiman, 2001) to predict log population density. Random Forest (RF) models are an ensemble, nonparametric modeling approach that grows a "forest" of individual classification or regression trees and improves upon bagging (Breiman, 1996) by using the best of a random selection of predictors at each node in each tree (Breiman, 2001; Liaw & Wiener, 2002).

As expected when combining multiple observations that are mostly independent, the best, most unbiased prediction was arrived at by taking the mean of all trees within the forest and back-transforming the log to arrive at an estimate of per-pixel population density. Medians and percentile ranges were also assessed as alternative approaches for prediction; however, the back-transformed mean consistently out-performed the alternative summary methods during validation. The resulting country-wise population density map was then used as a weighting layer for a standard dasymetric mapping approach as described for the AfriPop and AsiaPop (now WorldPop) data sets by (Gaughan, et al., 2013; Linard, et al., 2012; Linard & Tatem, 2012; Tatem et al., 2007).

2.3.3 Accuracy Assessment

The four output population maps produced using administrative level 1 input census data (Fig. 1), were then compared to the level 2 census counts to provide one method of assessing mapping accuracies, following Gaughan et al. (2013). The individual cell values of the output population maps represent people per cell, and were then added together for each census unit. These “predicted” sums were then compared with the observed census counts within each unit. Summary statistics were then calculated, including root mean square error (RMSE), the RMSE divided by the mean census unit count (%RMSE) and the mean absolute error (MAE). Together these statistics were used to compare the predictive ability of each methodology.

3. Results

3.1 Urban Extraction Results

For the urban extraction results in the test areas, all of them referring to Landsat scenes recorded in 2007, the validation was performed as follows: human settlement extents were manually digitized from Very High Resolution (VHR) Quickbird images available in Google Earth™, and recoded in 2007, if possible in the same month of the corresponding Landsat scene. The relatively small cities of Manado and Bandung, as well as the big urban agglomeration of Jakarta were considered.

The mapping results are shown in Fig. 4, while the quantitative validation results for Manado with and without spatial post-processing (see section 2.2.2) are reported in Table 2. Visually, the approach shows an accurate extraction of the human settlement extents at the pixel level, with a few misclassifications outside the actual urban area, and missing areas within the boundary of the larger blocks. The quantitative evaluation shows instead a large omission error percentage. After post-processing, however, the overall accuracy improves to 85% and the omission error decreases from 87% to below 19%. Satisfied with the relative accuracy of detecting urban areas using the NDSV classifier on the GEE system, the process was applied to three collections on the Google Earth Engine, and then integrated with the MDA Landcover dataset. This combined land cover dataset, using the GEE-derived built area delineations was then applied to the population mapping process and evaluated statistically for prediction accuracy.

A small sample of the urban extents generated for tests 2, 3 and 4 are shown in Fig. 5 for the central part of Jakarta along with the urban extents for the same area in the MDA data set.

3.2 Random Forest Statistical Output

The differences between results are determined by the ancillary datasets used in the population mapping detailed in Table 1.

Referring to the covariate names in Table 1, there are two significant covariates in the Random Forest mapping process, “BLT” (Built) and “lig” (VIIRS Nightlights). Table 3 provides some insight into the importance of the variables in the mapping process by showing how much Mean Squared Error (MSE) increases when the specified covariate is randomly permuted and predictions re-calculated. The most important variables include the “BLT” covariates, indicating “Built” areas, which include urban and rural settlements. In addition, for all tests, except for Test 4 (GEE 2008-2010), the “lig” (VIIRS Nightlights data) have higher importance than other covariates.

Table 3 also displays the increase in node purity in each test, which documents reduction in residual sum of squared error for the predictions at the ends of the branches of each tree when the specified variable is used during the Random Forest mapping process. Again referring to the variables detailed in Table 1, we show that the “BLT” (built) classes with the GEE integrations in Tests 2, 3 and 4 are the most important in the Random Forest process.

Again referring to the variables detailed in Table 1, it can be observed that the “BLT” classes with the GEE integrations in Tests 2, 3 and 4 are making the built classes the most important in the Random Forest process.

3.3 Random Forest Accuracy Assessment

The accuracy assessment process detailed in Section 2.3.3 shows how much the urban extents improve the output when the census data were aggregated from district to province. The

tests in the previous sections detail how well the RF does in predicting population values at the census unit level, but more importantly is whether the population map produced using built land cover data from the three GEE-derived approaches is better at redistributing the population numbers from coarser census units. Two different error assessment methods are presented: root mean square error (RMSE), also expressed as a percentage of the mean population size of the administrative level (% RMSE); and the mean absolute error (MAE).

For both RMSE and MAE, the results in Table 4 indicate that Test 4 increased population mapping accuracy the most, with Test 3 slightly better than Test 1. Notably, the urban extraction from Test 2, which used built extents derived from years 2006 to 2008 had the lowest redistribution accuracy. It is notable that the landcover changes allow for Test 3 and Test 4 to outperform the MDA dataset in reducing error, creating more concurrent built data to correlate better with our other datasets.

4. Discussion and conclusions

The possibilities that the Google Earth Engine offers in analyzing remotely sensed data on a global scale with the power of Google's cloud computing are substantial. The inclusion of continuously updated Landsat data along with classification tools and significant processing power will enable newer and more accurate ways to map human settlements across large areas at 30 m spatial resolution, document past changes and continually update current estimates. The potential of this resource has been recently illustrated for multitemporal forest mapping (Hansen et al., 2013), and here we outline initial steps for similar efforts in human settlement and population mapping.

The application of the NDSV within the GEE shows significant potential for settlement mapping within the tool. Characterizing human settlements can be considered as a binary problem, but where the “non-urban” class is very heterogeneous. It therefore requires a classifier which is non-parametric, i.e. that does not assume any peculiar statistical distributions of the input values. Moreover, since the NDSV is built through a composition of 15 bands, the classifier has to be able to manage high-dimensional spaces. Therefore, classifiers developed for hyperspectral data are preferable, using, for example, the spectral angle mapper classifier (Angiuli & Trianni, 2014), that captures the differences in multispectral vectors and is robust with respect to difference in illumination. Since this classifier is not available in the GEE environment, Support Vector Machines (SVM) and Classification and Regression Trees (CART) were considered instead ("Earthengine-api - Earth Engine Access Library - Google Project Hosting," 2014), with similarly strong results shown.

Both the SVM and CART are suitable to binary problems, but our tests suggested that CART produced more accurate urban extent maps. The statistical indices explored in the Random Forest population mapping process in Table 3 highlight to what degree the distance to “built” environments (lan_dstBLT) covariate plays a role in reducing error and increasing the quality of the output of the population mapping process. When the focus was on which variable, if removed, would increase the RMSE, the GEE experiments (Tests 2, 3, and 4) showed that the distance to “built” covariate was an important one. Table 3 also reflects the same results in increasing node purity in the process. It is important to note that in tests 2 and 4, urban extents extracted in 3 consecutive years are combined, while in test 3 a single year is considered. Test 2 showed the greatest amount of error, utilizing urban extents that were obtained from the GEE for years 2006, 2007 and 2008. It is clear that the modification of the landcover from test 1 for the

same time period reflected in test 2, changes the areas within Jakarta significantly. The improved accuracy of test 1 over test 2 could just reflect a better correlation of values instead of informing what is making the data more spatially significant, and in that circumstance, it can be argued that the GEE urban extents can be a critical component in the creation of multitemporal datasets that can modify existing landcover datasets in order to examine trends, in an efficient manner, for different years. Table 4 shows how the integration of GEE extents correlates well in the population mapping process and decreases error, by adding more concurrent built data along with our other covariate datasets.

In using census data from 2010, land cover data closest to this year stands a better chance of being the best proxy for disaggregation if all other factors are equal. In this sense, there is an inherent bias in the tests, but it also highlights the benefits of the GEE approach, that is being able to produce an accurate urban extent map for any time period, with the ability to match up land cover data to particular census dates.

Overall, the NDSV is shown here to be a reliable method to detect urban extents, especially when using a powerful tool to analyze the data such as the GEE. Moreover, the GEE represents one of the most powerful tools offered today in remote sensing with its ability to analyze and classify remotely sensed data over different temporal scales. Finally, the use of NDSV derived extents produced in the GEE and integrated in a flexible population mapping method enables testing of the validity of the classifications in improving population distribution mapping, providing an additional novel accuracy assessment approach. As urbanization processes continue to accelerate in many countries around the world, accurate, powerful and efficient methods for rapid mapping of settlements and their changes, as well as populations within them are a prerequisite for strategic planning and impact assessments. The results here

point towards the integration of classification and population mapping methods within GEE as a way of meeting this need.

Acknowledgements:

The work by NNP during his stay at the University of Pavia was supported by the CARIPLO Foundation, Project 2009-2936 for the Internationalization of the PhD programs in ICT and Bioengineering of the University of Pavia. PG acknowledges support from the Google Foundation for the preliminary phase of this research. AJT acknowledges funding support from the RAPIDD program of the Science and Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health, and is also supported by grants from the Bill and Melinda Gates Foundation (#49446, #1032350). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This work forms an output of the WorldPop project (www.worldpop.org.uk) and Flowminder Foundation (www.flowminder.org).

References:

Alberti, M., Weeks, R., & Coe, S. (2004). Urban Land-Cover Change Analysis in Central Puget Sound. *Photogrammetric Engineering & Remote Sensing, 70*(9), 1043-1052. doi: 10.14358/PERS.70.9.1043

Angel, S., Sheppard, S. C., Civco, D. L., Buckley, R., Chabaeva, A., Gitlin, L., Kraley, A., Parent, J., Perlin, M. (2005). The dynamics of global urban expansion. Washington, D.C.: World Bank, Transport and Urban Development Department.

Angiuli, E., & Trianni, G. (2014). Urban Mapping in Landsat Images Based on Normalized Difference Spectral Vector. *IEEE Geoscience and Remote Sensing Letters, 11*(3), 661-665. doi: 10.1109/LGRS.2013.2274327

Bagan, H., & Yamagata, Y. (2012). Landsat analysis of urban growth: How Tokyo became the world's largest megacity during the last 40 years. *Remote Sensing of Environment, 127*, 210-222. doi: 10.1016/j.rse.2012.09.011

Bhaduri, B., Bright, E., Coleman, P., & Urban, M. L. (2007). LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal, 69*(1-2), 103-117. doi: 10.1007/s10708-007-9105-9

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

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. doi:10.1023/A:1010933404324

Burchfield, M., Overman, H. G., Puga, D., & Turner, M. A. (2006). Causes of Sprawl: A Portrait from Space. The Quarterly Journal of Economics, 121(2), 587-633. doi: 10.1162/qjec.2006.121.2.587

Carpenter, D., & Carpenter, S. (1983). Modeling inland water quality using Landsat data.*Remote Sensing of Environment, 13*(4), 345-352. doi: 10.1016/0034-4257(83)90035-4

Center for International Earth Science Information Network (CIESIN), Columbia University, International Food Policy Research Institute (IFPRI), The World Bank and Centro International de Agricultura Tropical (CIAT). (2004). Global Rural–Urban Mapping Project (GRUMP): Urban Extents. Palisades, New York, CIESIN, Columbia University.

Earthengine-api - Earth Engine Access Library - Google Project Hosting. (2014). Retrieved February 1, 2014, from http://code.google.com/p/earthengine-api/

Elvidge, C., Baugh, K., Kihn, E. A., & Davis, E. R. (1996). Mapping city lights with nighttime data from the DMSP operational linescan system. Photogrammetric Engineering and Remote Sensing, 63, 727 – 734.

Elvidge, C. D., Baugh, K. E., Dietz, J. B., Bland, T., Sutton, P. C., & Kroehl, H. W. (1999). Radiance calibration of DMSP-OLS low-light imaging data of human settlements. Remote Sensing of Environment, 68, 77 – 88.

Elvidge, C., Baugh, K. E., Hobson, V. R., Kihn, E. A., Kroehl, H. W., Davis, E. R., et al. (1997). Satellite inventory of human settlements using nocturnal radiation emissions: A contribution to the global toolchest. Global Change Biology, 3, 387 – 395.

Foody, G. M. (2000). Estimation of sub-pixel land cover composition in the presence of untrained classes. Computers & Geosciences, 26, 469 – 478.

Gaughan, A. E., Stevens, F. R., Linard, C., Jia, P., & Tatem, A. J. (2013). High Resolution Population Distribution Maps for Southeast Asia in 2010 and 2015. (F. Pappalardo, Ed.) PLoS ONE, 8(2), e55882. doi:10.1371/journal.pone.0055882

GeoHive. (2014). - *Population Statistics*. Retrieved October 1st, 2013, from http://www.geohive.com/

Global Administrative Areas. (2014). *Global Administrative Areas*. Retrieved October 1st, 2013, from http://www.gadm.org/

Guindon, B., Zhang, Y., & Dillabaugh, C. (2004). Landsat urban mapping based on a combined spectral–spatial methodology. Remote Sensing of Environment, 92(2), 218-232. doi: 10.1016/j.rse.2004.06.015

Hansen, M. C., P. V. Potapov, R. Moore, M. Hancher, S. A. Turubanova, A. Tyukavina, D. Thau, S. V. Stehman, S. J. Goetz, T. R. Loveland, A. Kommareddy, A. Egorov, L. Chini, C. O. Justice, & J. R. G. Townshend (2013). High-Resolution Global Maps of 21st-Century Forest Cover Change. Science, 342(6160), 850-853. doi: 10.1126/science.1244693

Hansen, M., DeFries, R., Townshend, J. R. G., & Sohlberg, R. (1998).1km land cover classification derived from AVHRR. College Park, Maryland. The Global Land Cover Facility.

IUCN and UNEP. (2012). The World Database on Protected Areas (WDPA). Cambridge, UK: UNEP-WCMC. Retrieved October 15th, 2013 from http://www.protectedplanet.net

Lehner, B., Verdin, K., Jarvis, A., & Fund, W. W. (2006). HydroSHEDS Technical Documentation (p. 27). World Wildlife Fund.

Liaw, A., & Wiener, M. (2002). Classification and Regression by randomForest. R News, 2(3), 18–22.

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., & Tatem, A. J. (2012). Population distribution, settlement patterns and accessibility across Africa in 2010. PloS ONE, 7(2), e31743. doi:10.1371/journal.pone.0031743

Linard, C., Gilbert, M., & Tatem, A. J. (2011). Assessing the use of global land cover data for guiding large area population distribution modelling. GeoJournal, 76(5), 525-538. doi: 10.1007/s10708-010-9364-8

Linard, C., & Tatem, A. J. (2012). Large-scale spatial population databases in infectious disease research. International Journal of Health Geographics, 11, 7. doi:10.1186/1476-072X-11-7

MDA Federal Inc. (2007). EarthSat GeoCover LC Overview. Retrieved October 15, 2013, from http://www.mdafederal.com/geocover/geocoverlc/gclcoverview

NGA. (2005). Vector Map (VMap) Level 0. National Geospatial-Intelligence Agency (NGA). Retrieved April 24, 2013, from http://geoengine.nga.mil/geospatial/SW_TOOLS/NIMAMUSE/webinter/rast_roam.html

NOAA. (2012). VIIRS Nighttime Lights - 2012. Earth Observation Group, National Geophysical Data Center, National Oceanic and Atmospheric Administration (NOAA). Retrieved April 08, 2013, from http://www.ngdc.noaa.gov/dmsp/data/viirs_fire/viirs_html/viirs_ntl.html

OSM. (2013). OpenStreetMap Base Data. OpenStreetMap.org. Retrieved October 15th, 2013, from http://www.openstreetmap.org/

Pesaresi, M., Gerhardinger, A., & Kayitakire, F. (2008). A Robust Built-Up Area Presence Index by Anisotropic Rotation-Invariant Textural Measure. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 1(3), 180-192. doi: 10.1109/JSTARS.2008.2002869

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

Rawashdeh, S. A., & Saleh, B. (2006). Satellite Monitoring of Urban Spatial Growth in the Amman Area, Jordan. *Journal of Urban Planning and Development, 132*(4), 211. doi: 10.1061/(ASCE)0733-9488(2006)132:4(211)

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

Schneider, A. (2012). Monitoring land cover change in urban and peri-urban areas using dense time stacks of Landsat satellite data and a data mining approach. Remote Sensing of Environment, 124, 689-704. doi: 10.1016/j.rse.2012.06.006

Schneider, A., Friedl, M. A., & Potere, D. (2009). A new map of global urban extent from MODIS satellite data. *Environmental Research Letters, 4*(4), 044003. doi: 10.1088/1748-9326/4/4/044003

Schneider, A., Friedl, M. A., & Potere, D. (2010). Mapping global urban areas using MODIS 500-m data: New methods and datasets based on 'urban ecoregions'. *Remote Sensing of Environment, 114*(8), 1733-1746. doi: 10.1016/j.rse.2010.03.003

Schneider, A., Friedl, M. A., McIver, D. K., & Woodcock, C. E. (2003). Mapping urban areas by fusing multiple sources of coarse resolution remotely sensed data. Photogrammetric Engineering and Remote Sensing, 69, 1377 – 1386

Sexton, J. O., Song, X., Huang, C., Channan, S., Baker, M. E., & Townshend, J. R. (2013). Urban growth of the Washington, D.C.–Baltimore, MD metropolitan region from 1984 to 2010 by annual, Landsat-based estimates of impervious cover. Remote Sensing of Environment, 129, 42-53. doi: 10.1016/j.rse.2012.10.025

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2014). Disaggregating census data for population mapping using Random Forests with remotely-sensed and other ancillary data. *Manuscript Submitted for Publication*.

Sutton, P. (2003). A scale-adjusted measure of urban sprawl using nighttime satellite imagery. Remote Sensing of Environment, 86, 353 – 369.

Tatem, A. J., Noor, A. M., Von Hagen, C., Di Gregorio, A., & Hay, S. I. (2007). High resolution population maps for low income nations: combining land cover and census in East Africa. PloS ONE, 2(12), e1298. doi:10.1371/journal.pone.0001298

Taubenböck, H., Esch, T., Felbier, A., Wiesner, M., Roth, A., & Dech, S. (2012). Monitoring urbanization in mega cities from space. Remote Sensing of Environment, 117, 162-176. doi: 10.1016/j.rse.2011.09.015

The WorldPop Project. (2014). Retrieved from http://www.worldpop.org.uk/

Woodcock, C.E., Allen, R., Anderson, M., Belward, A., Bindschadler, R., Cohen, W., Gao, F., Goward, S.N., Helder, D., Helmer, E., Nemani, R., Oreopoulos, L., Schott, J., Thenkabail, P.S., Vermote, E.F., Vogelmann, J., Wulder, M.A., Wynne, R. (2008). Free Access to Landsat Imagery. *Science, 320*(5879), 1011a-1011a. doi: 10.1126/science.320.5879.1011a

Xu, H. (2008). A new index for delineating built-up land features in satellite imagery. *International Journal of Remote Sensing, 29*(14), 4269-4276. doi: 10.1080/01431160802039957

Yuan, F., Sawaya, K. E., Loeffelholz, B. C., & Bauer, M. E. (2005). Land cover classification and change analysis of the Twin Cities (Minnesota) Metropolitan Area by multitemporal Landsat

remote sensing. *Remote Sensing of Environment, 98*(2-3), 317-328. doi: 10.1016/j.rse.2005.08.006

Tables

Table 1: Test-specific data sources and variable names used for population density estimation used for dasymetric weights.

Type	Variable Name(s)*	Description	Indonesia Java Data
Census		Country-specific census and scale	2010, Admin-level 2 (GADM,2014), (Geohive, 2014)
Land Cover	lan_cls011, lan_dst011	Cultivated terrestrial lands	Landcover Experiments detailed in Table 2
	lan_cls040, lan_dst040	Woody / Trees	
	lan_cls130, lan_dst130	Shrubs	
	lan_cls140, lan_dst140	Herbaceous	
	lan_cls150, lan_dst150	Other terrestrial vegetation	
	lan_cls160, lan_dst160	Aquatic vegetation	
	lan_cls190, lan_dst190	Urban area	
	lan_cls200, lan_dst200	Bare areas	
	lan_cls210, lan_dst210	Water bodies	
	lan_cls230, lan_dst230	No data, cloud/shadow	
	lan_cls240, lan_dst240	Rural settlement	
	lan_cls250, lan_dst250	Industrial area	
	lan_clsBLT, lan_dstBLT	Built, merged urban/rural class	
Continuous Raster-Format			
	lig	Lights at night	Suomi VIIRS-Derived

			(NOAA, 2012)
	tem	Mean temperature, 1950-2000	WorldClim/BioClim
	pre	Mean precipitation, 1950-2000	WorldClim/BioClim
	ele	Elevation	HydroSHEDS (Lehner et al., 2006)
	ele_slope	Slope	HydroSHEDS-Derived (Lehner et al., 2006)
Converted Vector-Format	roa_dst	Distance to roads	OSM (2013)
	riv_dst	Distance to rivers/streams	OSM (2013)
	pop_cls, pop_dst	Generic populated places	VMAP0 *merged*†
	wat_cls, wat_dst	Water bodies	World Food Programme
	pro_cls, pro_dst	Protected areas	WDPA, IUCN (2012)
	poi_cls, poi_dst	Populated Points	OSM (2013)
	bui_cls, bui_dst	Buildings	OSM (2013)

* The variable names are used in Random Forest model output and throughout the text as reference to the specific data they were derived from. The first three letters are derived from the data type (e.g. "lan" indicates land cover) and the last three letters, if present, indicates what type of data each variable represents (e.g. "_cls" is a binary classification and "_dst" is a calculated Euclidean distance-to variable.

† The default data for populated places is merged from several VMAP0 data sources. There are three VMAP0 data sets used: The point data pop/builtupp and pop/mispopp are buffered to 100 m and merged with the pop/builtupa polygons creating a vector-based built layer. This layer is then converted to binary class and distance-to rasters for use in modeling. (NGA, 2005)

Table 2. Confusion Matrices for Kota Manado without (top) and with (bottom) the Spatial Post-Processing Step (Fig. 4):

Overall Accuracy =(2242/4000) 56.05%			
	Ground Truth (Pixels)		
Class	urban	non urban	Total
urban	245	3	248
non urban	1755	1997	3752
Total	2000	2000	4000
Overall Accuracy = (3398/4000) 84.95%			
	Ground Truth (Pixels)		
Class	urban	non urban	Total
urban	1625	227	1852
non urban	375	1773	2148
Total	2000	2000	4000

Table 3: Top Five Statistical Outputs: Percent Increase of Mean Squared Error When Variable is Randomly Permuted and Total Decrease in Residual Sum of Squares When Variable is Selected For Decision Tree Node

Percent Increase of Mean Squared Error When Variable Randomly Permuted

Test 1 (MDA) (total of 23 covariates used), (81% Variance Explained)	**Test 2 (GEE 2006-2008) (total of 22 covariates used), (83% Variance Explained)**
20.2 (Lights)	18.4 (Lights)
12.7 (Landcover Distance to Built Areas)	17.4 (Landcover Distance to Built Areas)
10.8 (Distance to Populated Points)	10.6 (Distance to Populated Points)
9.24 (Distance to Buildings)	7.91 (Distance to Generic Population Places, VMAP0)
9.08 (Landcover Distance to Cultivated Terrestrial Areas)	7.79 (Landcover Distance to Cultivated Terrestrial Areas)
Test 3 (GEE 2009) (total of 22 covariates used), (83% Variance Explained)	**Test 4 (GEE 2008-2010) (total of 23 covariates used), (84% Variance Explained)**
19.3 (Lights)	19.8 (Landcover Distance to Built Areas)
16.5 (Landcover Distance to Built Areas)	18.8 (Lights)
9.13 (Distance to Populated Points)	8.00 (Distance to Populated Points)
7.13 (Distance to Roads)	7.77 (Landcover Distance to Cultivated Terrestrial Areas)
6.90 (Landcover Distance to Cultivated Terrestrial Areas)	7.23 (Distance to Roads)
Total Decrease in Residual Sum of Squares When Covariate Used	
Test 1 (MDA) (total of 23 covariates used) , (81% Variance Explained)	**Test 2 (GEE 2006-2008) (total of 22 covariates used), (83% Variance Explained)**
53.8 (Lights)	49.3 (Landcover Distance to Built Areas)
31.7 (Landcover Distance to Built Areas)	43.7 (Lights)
19.1 (Distance to Roads)	17.7 (Distance to Populated Points)
17.2 (Distance to Populated Points)	16.9 (Distance to Roads)
16.9 (Distance to Buildings)	13.7(Distance to Buildings)
Test 3 (GEE 2009) (total of 22 covariates used), (83% Variance Explained)	**Test 4 (GEE 2008-2010) (total of 23 covariates used), (84% Variance Explained)**
56.5 (Landcover Distance to Built Areas)	58.4 (Landcover Distance to Built Areas)
46.6 (Lights)	45.3 (Lights)
15.3 (Distance to Roads)	15.3 (Distance to Roads)
13.5(Distance to Populated Points)	12.0 (Distance to Populated Points)
9.76 (Distance to Generic Populated Places, VMAP0)	11.2 (Distance to Buildings)

Table 4: Accuracy Assessment Results for Four Urban Land Cover Treatments

	RMSE	%RMSE	MAE
Test 1 (MDA)	1450.286	0.129064	787.8362
Test 2 (GEE 2006-2008)	2277.501	0.20268	1352.685
Test 3 (GEE 2009)	1377.889	0.122621	773.6329
Test 4 (GEE 2008-2010)	1346.32	0.119812	759.3168

Figures

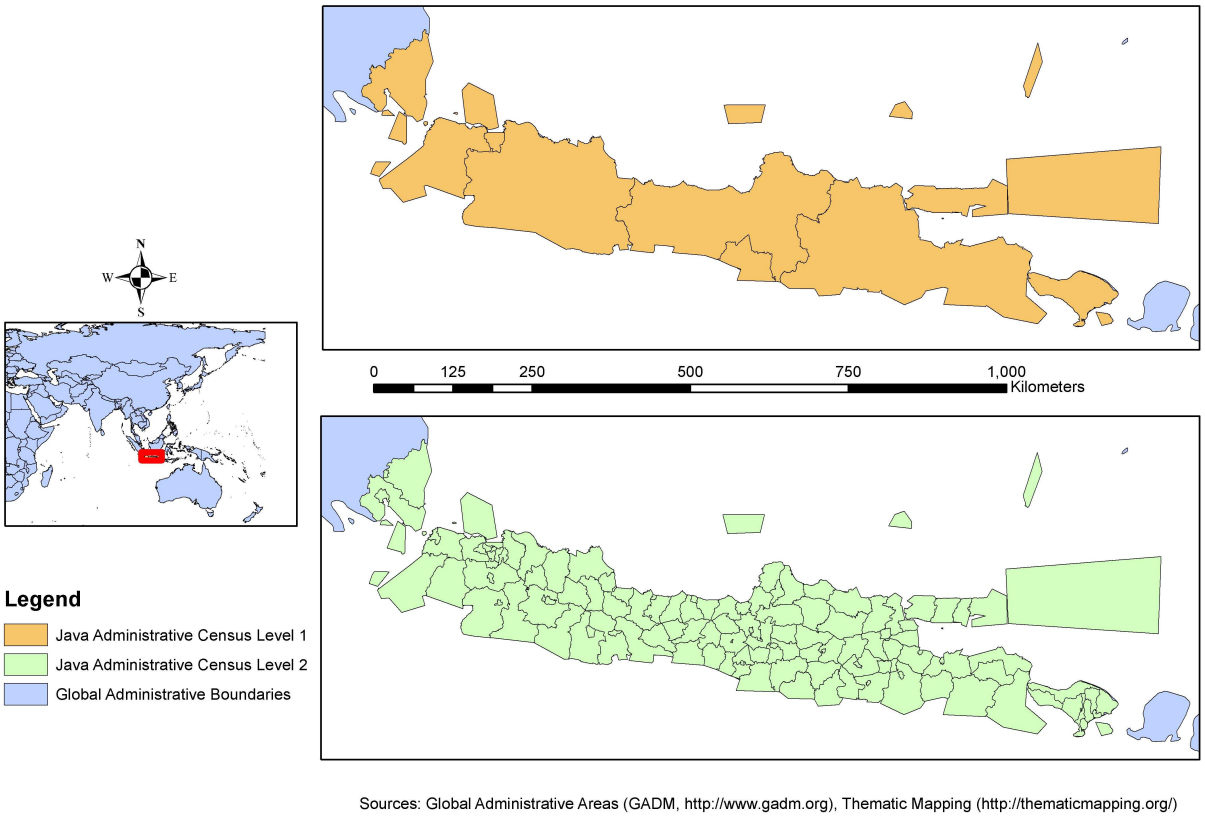

Fig. 1: Map of study area and Java administrative boundaries levels 1 and 2

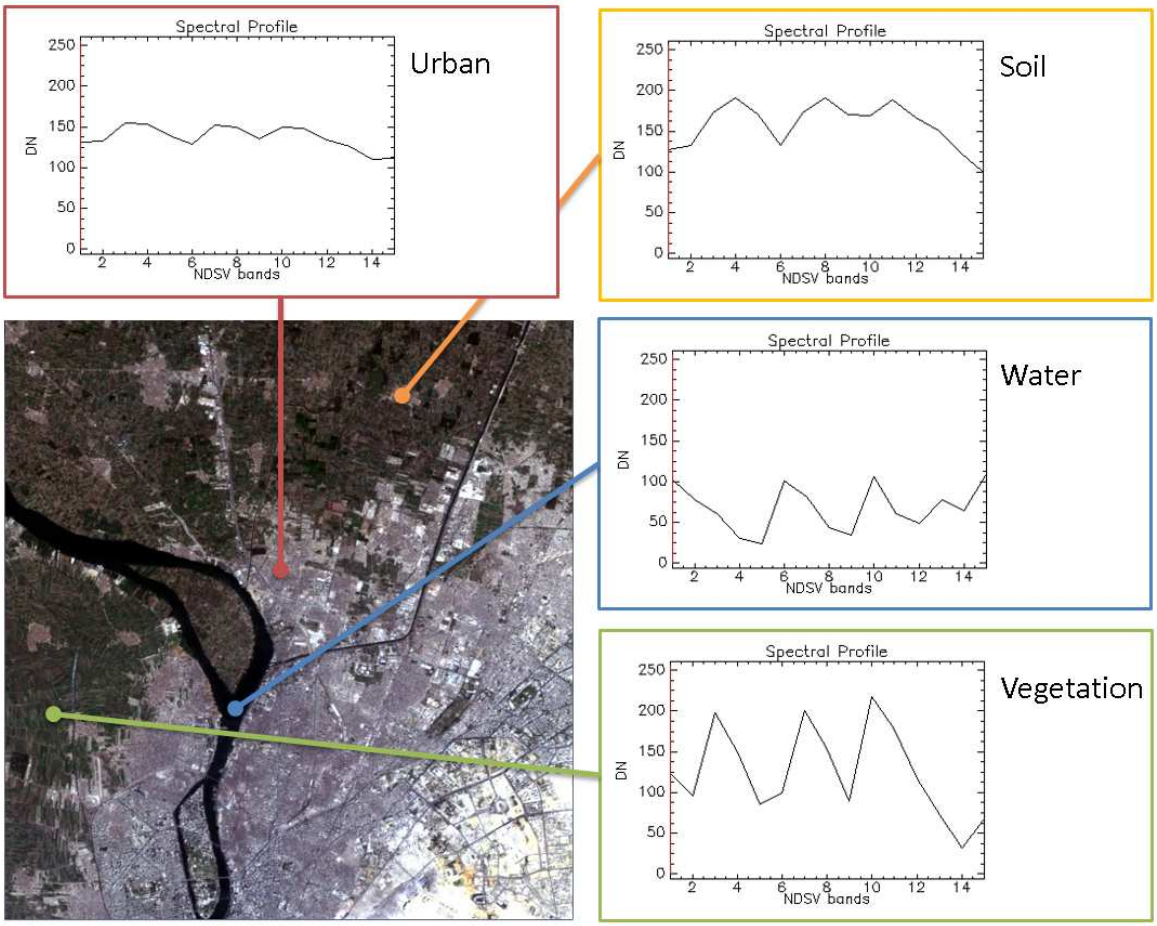

Fig. 2. Normalized Difference Spectral Vector (NDSV) profiles for urban areas, vegetation, water and bare soil.

Fig. 3: General structure of the data processing and map production procedure used to compare

the methodology outlined in Stevens et al (2014). The orange boxes represent items that are specific to the research presented here and not part of end-user map data product generation. The green boxes represent data pre-processing stages. Items in blue represent Random Forest model estimation, per-pixel prediction and dasymetric redistribution of census counts.

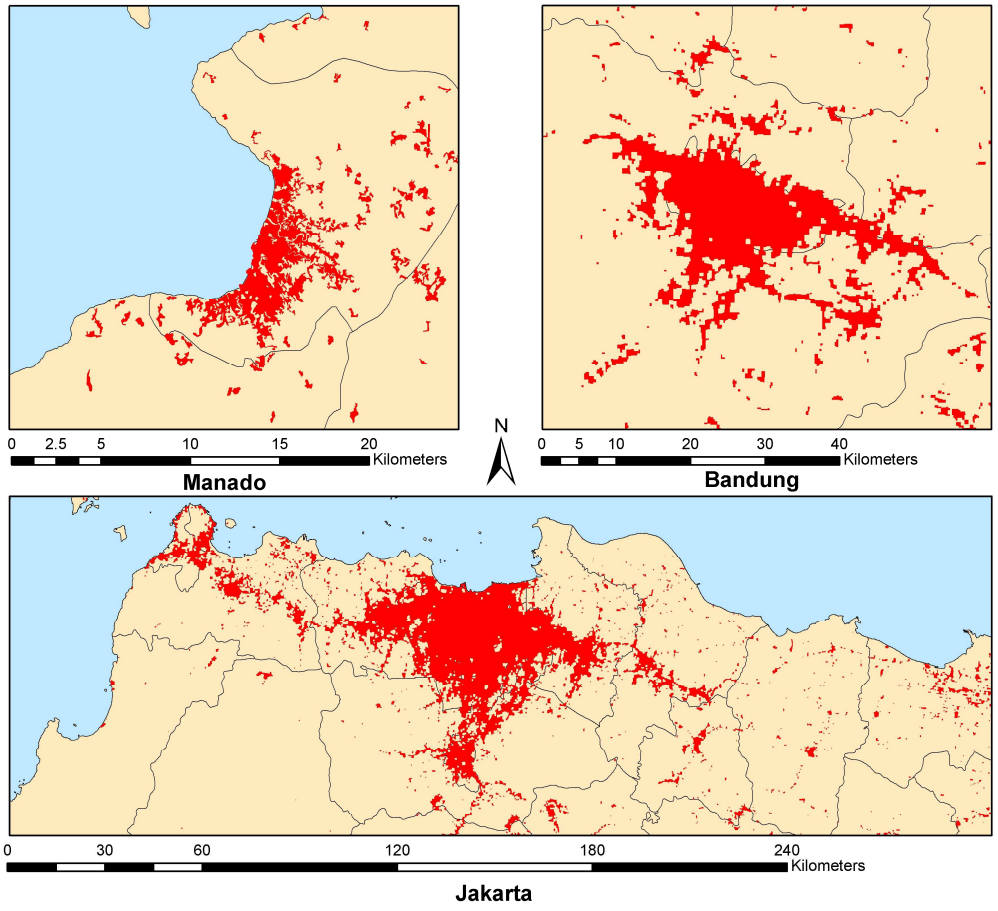

Fig. 4: Human settlement extraction results for Manado, Bandung and Jakarta, in Indonesia.

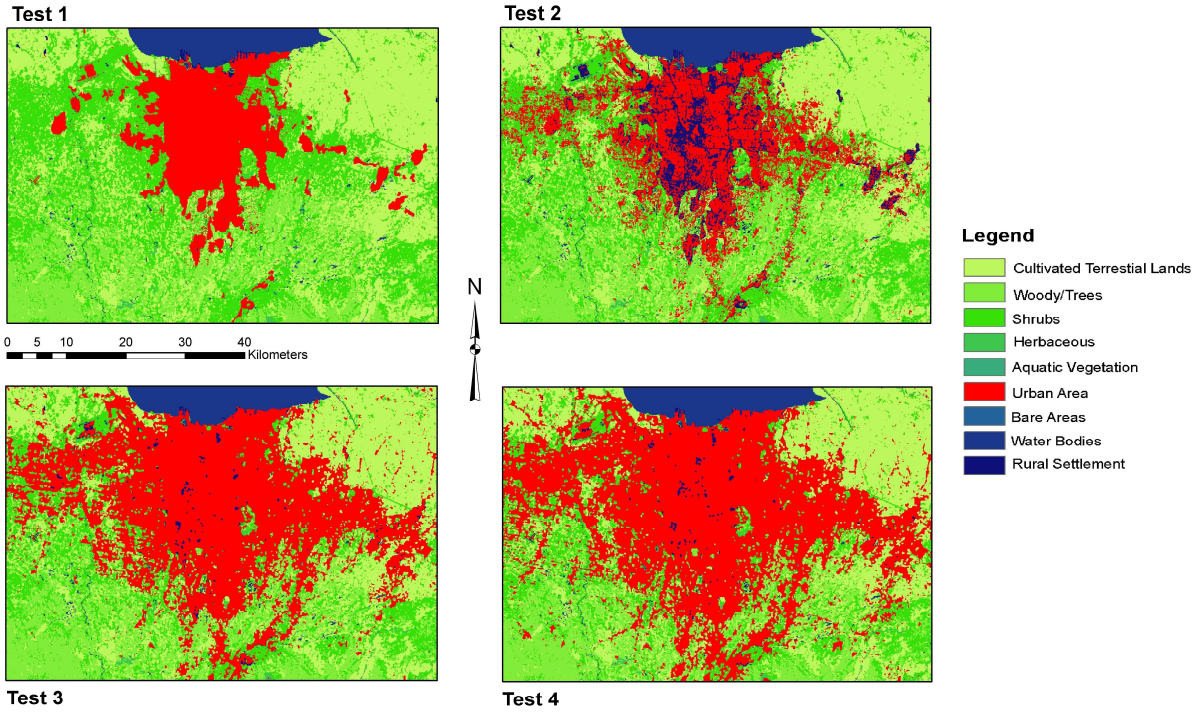

Fig. 5: A small sample of the area around central Jakarta used in the tests. Test 1 displays the EarthSat GeoCover Land Cover Thematic Mapper from MDA Federal (reflecting extents from 2007). Tests 2, 3 and 4 represent the Google Earth Engine derived extents that are merged into Test 1. Test 2 integrates urban areas from 3 collections from 2006, 2007 and 2008 (a collection for each year), Test 3 integrates urban areas from 3 collections in 2009, and Test 4 integrates urban areas from 2008, 2009, and 2010 (a collection for each year). Classifications reflected in the Legend are all from the MDA Federal dataset other than the “Urban Area” class, which was obtained from the Google Earth Engine derived NDSV extents.