Detecting Change in Urban Extent and Density: Machine Learning of Land Use Classes Applied to Houston, Texas

Gerardo Rodríguez Vázquez[1], sunglyoung Kim[1], Nicholas Jones[1], and Gaurav Bhardwaj[1]

[1]NYU Center for Urban Science & Progress

May 4, 2018

1. Introduction:

In this paper, we use machine learning to address a key challenge faced by the urban planning community: a shortage of consistent and usable information on land use patterns. Regulating land use is the raison d'être of urban planners. Through zoning regulations and long-range planning documents, they aim to shape urban growth towards favored patterns of spatial extent, density, and social or economic development. However, planners typically face significant data gaps. In New York, for instance, then city's Open Data Portal lacks any data on historical extent and density of the city. In this work, we utilize satellite imagery and machine learning to detect changes in urban extent and density. Using the rich availability of multi-spectral imagery accessible for free on Google Earth Engine, we train a classifier to detect key land use categories.

Using Houston, Texas, as our test case, we construct a Random Forest classifier. Following tuning and optimization of the classifier, we apply it to historical satellite imagery for the years 1999, 2003, 2007, 2011 and 2015. Our classifier successfully distinguishes urban extent from non-urban land and open water. It achieves encouraging levels of accuracy in distinguishing high-density and low-density urban areas. Applying these methods to the Greater Houston area, we identify those counties that underwent rapid rural-to-urban land conversion - such as the prosperous Highlands suburb. To further increase prediction accuracy, we prototype a method to use OpenStreetMap data alongside satellite imagery. Our results offer a method whereby planners can 'reality check' their intuitions about which parts of a city expanded or densified in recent decades.

2. Motivation and Literature Review

2.1 Scarce planning data; plentiful satellite data

Urban planners intervene in land use in several ways, whether to relieve traffic congestion through transport planning, preserve neighborhood character through residential density limitations, or separate industry from housing to preserve health (Hoch 2012). However, objective information on land use can be scarce. Geographic Information Systems (GIS) used by city planning departments are typically built on cadastral (ie. property tax) records combined with census-based demographic information (Landis 2012). Information on current and historical zoning information is frequently available, yet zoning maps may be disregarded in practice, and do not necessarily reflect the city's actual characteristics. Satellite data has attracted attention from urban researchers since the late 1970s given its potential to supplement existing urban planning data (Kontoes, 1999).

2.2 Machine learning for land use classification

The field of land use classification has expanded since the launch of the Landsat program in 1972. Landsat, a NASA-funded program, provides the longest-running consistent satellite imagery of the earth's surface (https://landsat.gsfc.nasa.gov/landsat-1/). The scientific literature based upon Landsat imagery expanded first in earth science and ecology(han). Researchers have particularly capitalized on the satellite's multi-spectral imagery, which captures (at present) eight bands - from longwave radiation, through the visible light spectrum, to shortwave infra-red. Particular advances in machine learning based on Landsat imagery exploited the Normalized Difference Vegetation Index (NDVI), which measures the difference between near infrared (which chlorophyll in vegetation strongly reflects) and red light (which vegetation absorbs) (Erener et al., 2012).

Exploiting NDVI has enabled researchers to gain high-frequency estimates of crop productivity to inform farming decisions, and to build early warning systems for deforestation in regions such as the Amazon (Michaelsen et al., 1994). As the field of land use classification through machine learning has become increasingly established, researchers have gravitated towards Random Forest as the algorithm of choice. Advantages highlighted in the literature include computational efficiency - which is greater than Support Vector Machines. Landsat has also been used alongside night-light data to predict income and poverty levels (Jean et al., 2016).

2.3 Applications to urban extent and density

Building upon land use classification studies in ecology and earth science, a growing research literature applies it to urbanization. Multi-spectral imagery is well-suited to detect urban built-up areas: although impervious surfaces lack the same distinctiveness of absorptive pattern on the visible and near-infrared wavelengths that makes vegetation easy to detect, increased reflectivity at the thermal imaging ends of the spectrum help to detect surfaces such as concrete and brick (Ward et al., 2000). Studies in cities such as Kolkata and Ho Chi Minh City have used time series of satellite imagery to track changes in urban extent over time (Goldblatt et al, 2016). These researchers used supervised classification methods.

A key challenge faced in this literature was to establish the training data required for a supervised classification exercise. Goldblatt et al addressed the challenge two ways: firstly by taking Ho Chi Minh City's property tax database and deriving a land use map from it; and secondly by hand-classifying a gridded map of the city's extent, pixel-by-pixel, with the categories "urban residential", "urban non-residential", and "non-urban." The first effort was abandoned as the city's land use database was deemed insufficiently true with regard to actual land utilization. The second method proved effective, albeit time-consuming. This method allowed researchers to train a classifier on the training image, where pixel values correspond to land use category, and to predict new pixel values, using the bands of Landsat's multiple spectrums as input values.

3. Methods

3.1 Reference Data on Land Use

In this research, we evaluated several methods to acquire reference data for land-use classification of urban extent and density in the United States. We initially constructed a land use map of New York City based upon the Department of City Planning's zoning shapefiles. In Geopandas, we reclassified all city areas from their detailed zoning code (eg. R4 for mid-density residential; P for park) into three categories: residential, urban non-residential, and park/non-urban. However, the classes were unsatisfactory because New York's zoning codes are not reliable indicators of actual land utilization, while the three categories were seen to have limited utility for planning decisions given the largely static city boundaries and high prevalence of mixed use.

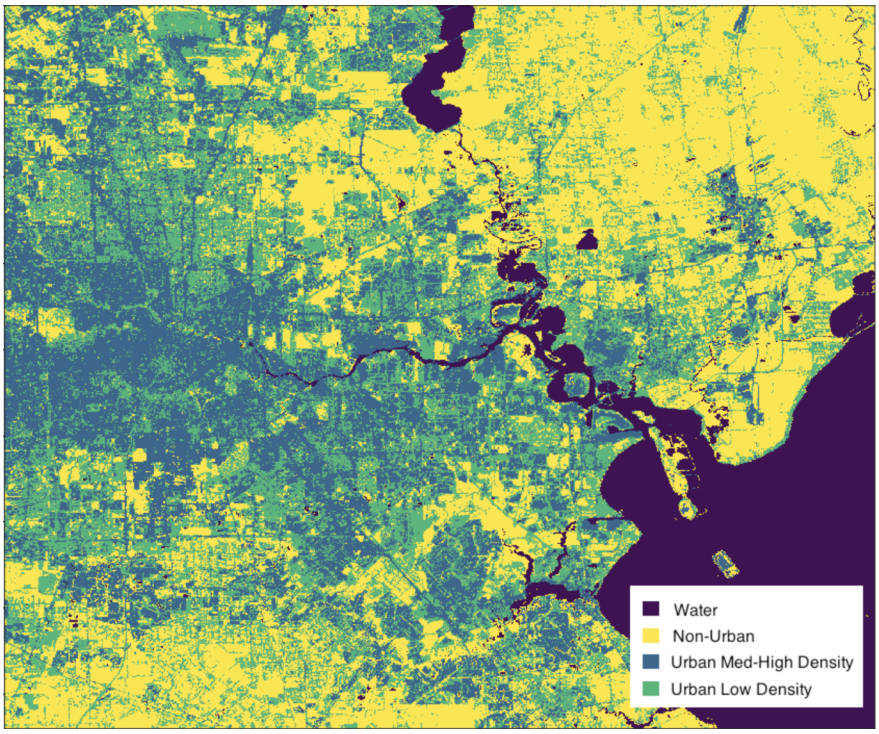

Figure 1: Land use in Houston in 2015: Reference image developed from Texas GIS Land Use Map

Having established that land use classification is of particular value in the context of fast-growing cities, we turned our attention to a list of the 10 fastest growing US cities, which are concentrated in sunbelt areas such as Texas, Nevada and Arizona. Other studies such as Goldblatt's (Goldblatt et al., 2018) have generated original training data for urban land use classification through hand-labeling of large raster files. In our research, we instead searched for existing detailed land-use maps in a fast-growing US region. A detailed land-use raster covering the greater Houston area for 2015 was acquired from the Texas GIS website. In Python, the pixel values were reclassified from the existing set of 10 categorical values (where 1-10 represented categories from open water through to dense urban areas, including varieties of non-urban land use such as wetlands and forest) to four categorical values: (1) open water; (2) urban: high density; (3) urban: low density; (4) non-urban land.

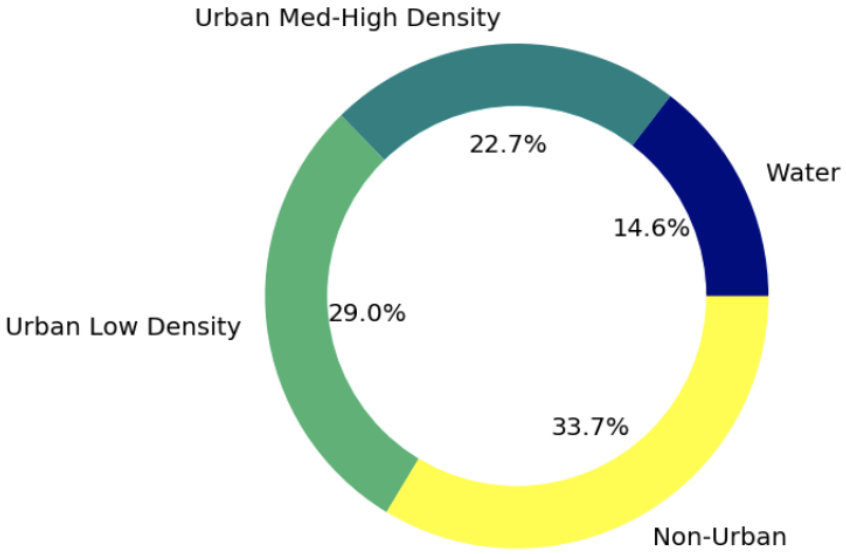

Figure 2: Land use classification scheme (by percentage share of region of interest, 2015)

3.2 Satellite Data

Using Google Earth Engine, we developed a script to preprocess and download Landsat-7 satellite imagery. Based on customizing existing script libraries, code was developed to: (i) collect one-year batches of satellite imagery; (ii) mask based on Landsat data's cloud-cover index band; (iii) select the non-cloudy pixels from the year's images; (iv) make a composite image from these pixels' median values; (v) export the image to Google Drive.

Images for 1999, 2003, 2007, 2011 and 2015 were created in this way. The imagery was downloaded at maximum resolution, with each pixel representing 30 square meters. Each image was converted into a Numpy array with dimensionality 2100 x 2528 x 11. Required image processing steps including clipping the reference image to the training image; removing NaN values; and resampling the reference image to achieve the exact same number of pixels and geographic area.

3.3 Random Forest classifier

Two classification methods were evaluated: Random Forest and Support Vector Machines. We proceeded with Random Forest having found a small advantage in classification accuracy and a substantial advantage in computation time - which was prohibitive in the case of SVM given our large data files.

A Random Forest was trained on the image data for 2015. Structuring the problem as a supervised classification exercise, we trained the Random Forest using the reference image as target value (or 'label' for each pixel) and the Landsat pixel values as the feature space.

Random Forest is an ensemble method that constructs k decision trees and (in the case of classification) takes the modal value of their output. Given these mechanics, we conducted a grid search to find key parameters

that would optimize the classifier, specifically (i) number of trees; (ii) maximum depth; and (iii) minimum sample leaf size. Accuracy was found to improve with increasing numbers of trees up to 12 but tail off after then. Given limited computational budget, the team proceeded with a 12-tree Random Forest classifier.

4. Model evaluation

As a proof of concept, we first trained the model on West Houston for 2015 and tested it for East Houston. We subsequently refined this approach by means of a k-folds cross-validation. The method is suited to our data, since we are able to train and test on the same 2015 dataset (the only year for which labeled data could be constructed). What the method does is, it divides the raster in 6 proportional cubes, trains the model on 5 of those cubes and tests it on the one left. This process is repeated for the number of cubes we have. We utilized a 6-fold cross-validation, since this is easily interpretable to audiences - for whom it can be visualized as testing a model on a one-sixth grid section of the city having learned it from the remaining 5/6ths of the image, and repeating until each pixel has been tested.

Overall accuracy of the classifier, measured by 6-fold cross-validation, was 76%.

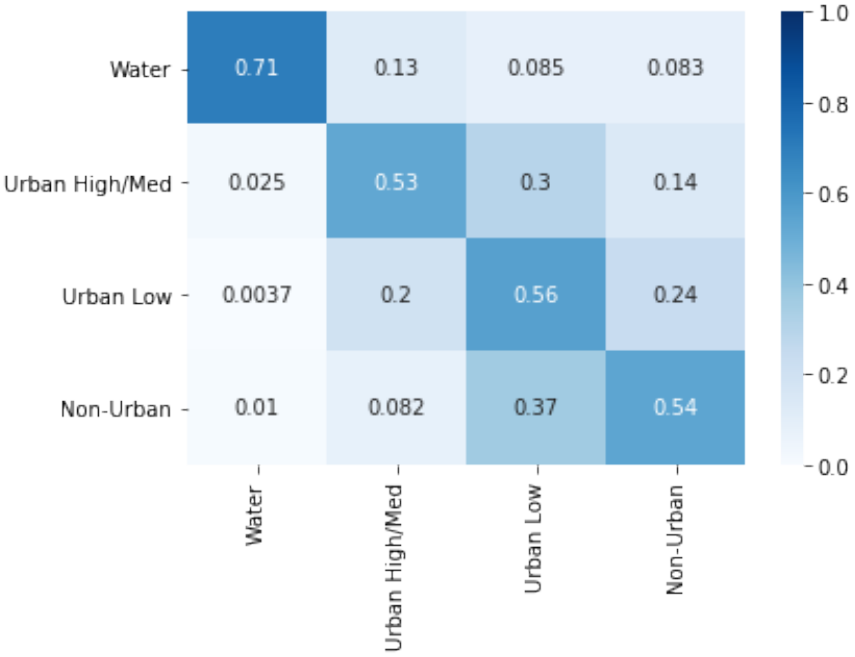

Figure 3: Land Use Predictions for Houston in 2015 vs Actual Values: True Positives Matrix

5. Findings and discussion

We built a machine learning classifier to detect changes in urban extent and density over time, and applied this to Houston, Texas. Having applied the classifier to satellite images from 1999 to 2015, we see a pattern

of expanding urban extent, particularly in the north-east suburbs. Areas such as the Highlands moved from non-urban to light urban, or light urban to dense urban.

The classifier achieves 76% accuracy based on a 6-fold cross-validation. However, error rates are higher between certain category pairs. Water is rarely mis-categorized, as expected given its distinctive reflective signature. However, dense urban is frequently miscategorized as light urban, while light urban is frequently miscategorized as dense urban or non-urban. Our classifier performs better if the two urban categories are grouped together, restricting the task to merely distinguishing built-up area from non-built up. Nevertheless, the findings are encouraging and demonstrate a proof-of-concept. Extensions aimed at raising classification performance further would include:

- Incorporating additional satellite imagery bands, such as Synthetic Aperture Radar imagery (an effective input for water detection) and MODIS thermal imagery (helpful to identify urban areas given their higher thermal reflectivity);
- Adding an NDVI layer through calculations based on the red and near-infrared Landsat bands; and
- Implementing a corner detection algorithm to pick up texture in urban environments, such as multiple dwelling roofs in dense urban areas.

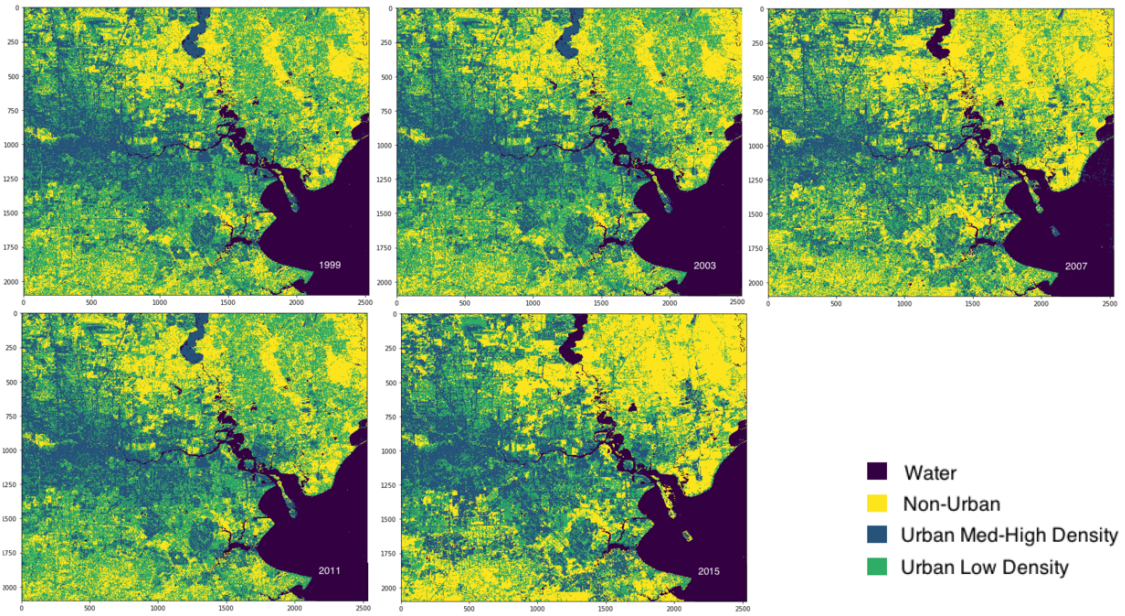

Figure 4: Output from Random Forest Classifier: Predicted Land Use in North-East Houston from 1999-2015

6. Proposed extension: OpenStreetMap feature density

As noted, the classifier currently mis-categorizes certain pairwise combinations of classes for which atmospheric reflectivity is similar. Distinguishing high-density urban areas from low-density urban areas is difficult, as is distinguishing low-density urban from non-urban land. We also note that in several images, such as the 1999 classifier output, water is mis-classified as urban land. To further improve the classifier's per-

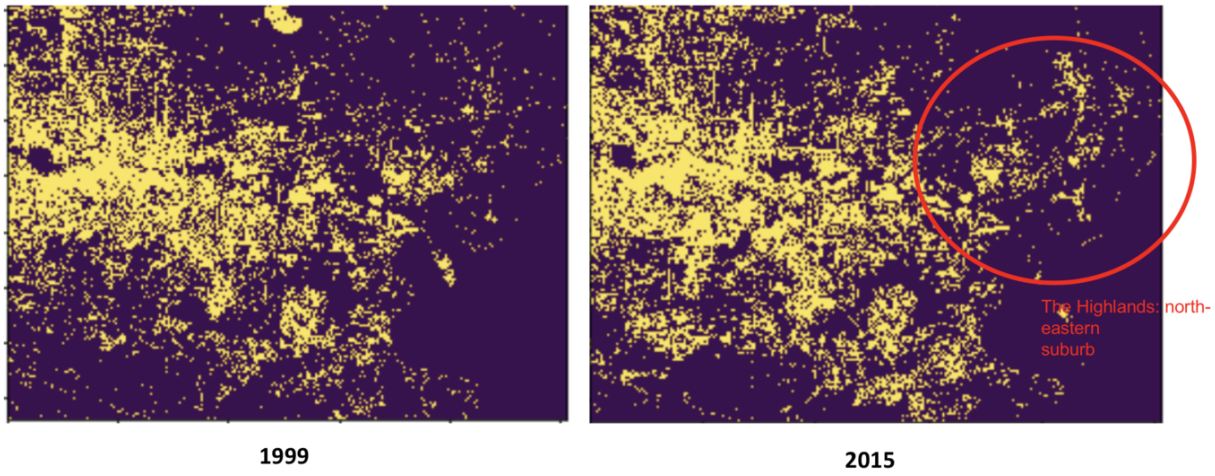

Figure 5: Output from Random Forest Classifier: Areas Classed as High Density Urban (North East Houston)

formance, we propose to use density of urban features shops, houses, intersections hospitals - as an addition the feature space.To demonstrate the production of these inputs, we produce raster images of urban features OpenStreetMap (OSM) tags. Accessible through OSM Application Programming Interface and its Overpass web interface, we acquire point locations of urban feature categories.

Kernel Density Estimation (KDE) is used produce raster images representing the mean density of the features. The KDE image, as indicated in Figure 5, can in future in incorporated into the model. In instances where our classifier mis-categorizes (for example) water as land, the KDE of OSM features is expected to improve accuracy, since the Random Forest decision trees can make additional cuts based upon the density of tags (captured in the raster pixel values from 0 - 255) or the lack of any tags on expanses of water.

OSM provides large numbers of tags. By incorporating this data source alongside Landsat's 7 bands (and any supplementary satellite imagery brought into the classifier in future), we anticipate improving the classification accuracy. Having expanded the feature space to incorporate OSM data on urban features, it is possible that the classifier might approach very high classification accuracy. If this is achieved, the new OSM data would make it possible to shift to more ambitious classification tasks. Based on constructing a new reference image, a classifier could be trained upon a larger number of categories, such as population densities per square feet or multiple land use categories including industrial, agricultural and commercial areas. Introducing the new data source of OSM - which has not been implemented in past studies so far as the researchers are aware - it may be possible to produce urban planning data of increasing detail and utility to the city planning community.

Conclusions

Random Forest classification offers an effective, low-cost way to detect changes in urban extent and density over time. We built a classifier that achieves 76% accuracy in a four-way classification task: distinguishing water, dense urban, light urban, and non-urban land. The method produces a time-series of classified areas which can be used to reality check city planners' understanding of their city's urban form: which areas grew fastest, and where density increased. Extending this method through additional input layers, or through engineered input layers such as NDVI and corner detection, could further enhance performance. The researched was possible due to the existence of detailed land use maps created by the Texas GIS services; without such reference data, the method is still feasible but may require creating hand-labeled reference

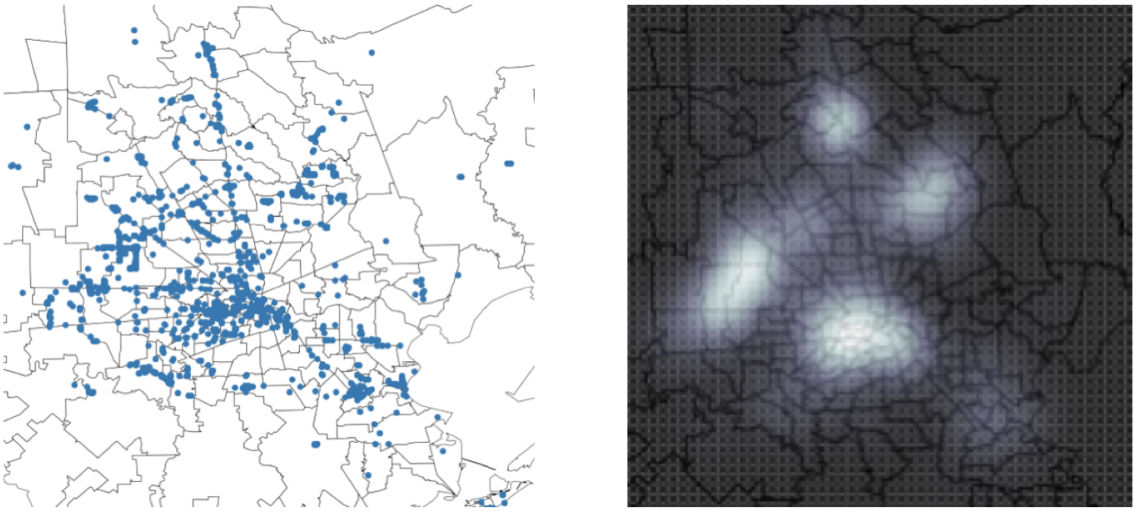

Figure 6: Latitude and longitude data of shops in Houston Texas, are extracted from Open Street Map and converted to KDE.

images based on property tax or zoning data. Such methods may be particularly valuable in fast-growing US cities - many concentrated in the Sunbelt - where urban planners are confronting rapid change with spatial patterns that are hard to apprehend; and in fast-growing cities of the developing world where urban planning data is sparse.

References

Landsat 7 Science Data Users Handbook. Landsat Handbook. URL https://landsat.gsfc.nasa.gov/wp-content/uploads/2016/08/Landsat7_Handbook.pdf. Accessed on Thu, May 03, 2018.

Arzu Erener, Sebnem Düzgün, and Ahmet Cevdet Yalciner. Evaluating land use/cover change with temporal satellite data and information systems. *Procedia Technology*, 1:385–389, 2012. doi: 10.1016/j.protcy.2012.02.079. URL https://doi.org/10.1016%2Fj.protcy.2012.02.079.

Ran Goldblatt, Klaus Deininger, and Gordon Hanson. Utilizing publicly available satellite data for urban research: Mapping built-up land cover and land use in Ho Chi Minh City Vietnam. *Development Engineering*, 3:83–99, 2018. doi: 10.1016/j.deveng.2018.03.001. URL https://doi.org/10.1016%2Fj.deveng.2018.03.001.

N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, aug 2016. doi: 10.1126/science.aaf7894. URL https://doi.org/10.1126%2Fscience.aaf7894.

Charalambos C. Kontoes. Image Analysis Techniques for Urban Land Use Classification. The Use of Kernel Based Approaches to Process Very High Resolution Satellite Imagery. In *Machine Vision and Advanced Image Processing in Remote Sensing*, pages 121–133. Springer Berlin Heidelberg, 1999. doi: 10.1007/978-3-642-60105-7_11. URL https://doi.org/10.1007%2F978-3-642-60105-7_11.

Joel Michaelsen, David S. Schimel, Mark A. Friedl, Frank W. Davis, and Ralph C. Dubayah. Regression Tree Analysis of satellite and terrain data to guide vegetation sampling and surveys. *Journal of Vegetation Science*, 5(5):673–686, oct 1994. doi: 10.2307/3235882. URL https://doi.org/10.2307%2F3235882.

Douglas Ward, Stuart R. Phinn, and Alan T. Murray. Monitoring Growth in Rapidly Urbanizing Areas Using Remotely Sensed Data. *The Professional Geographer*, 52(3):371–386, aug 2000. doi: 10.1111/0033-0124.00232. URL https://doi.org/10.1111%2F0033-0124.00232.