WORLD RESOURCES INSTITUTE

TECHNICAL NOTE

SPATIAL CHARACTERIZATION OF URBAN LAND USE THROUGH MACHINE LEARNING

PETER KERINS, EMILY NILSON, ERIC MACKRES, TAUFIQ RASHID, BROOKIE GUZDER-WILLIAMS, AND STEVEN BRUMBY

SUMMARY

This technical note describes the data sources and methodology underpinning a computer system for the automated generation of land use/land cover (LULC) maps of urban areas. Deploying a rich taxonomy to distinguish between different types of LULC within a built-up area, rather than merely distinguishing between artificial and natural land cover, enables a huge variety of potential applications for policy, planning, and research. Applying supervised machine learning techniques to satellite imagery yielded trained algorithms that can characterize LULC over a large spatial and temporal range, while avoiding many of the onerous constraints and expenses of the historical LULC mapping process: manual identification and classification of features. This note presents the construction and results of one such set of algorithms—city-specific convolutional neural networks—used to establish the technical viability of such an approach.

CONTENTS

Technical notes document the research or analytical methodology underpinning a publication, interactive application, or tool.

Suggested Citation: Kerins, P., E. Nilson, E. Mackres, T. Rashid, B. Guzder-Williams, and S. Brumby. 2020. "Spatial Characterization of Urban Land Use through Machine Learning." Technical Note. Washington, DC: World Resources Institute. Available online at: https://www.wri.org/publication/spatial-characterization-urban-land-use.

1. INTRODUCTION

1.1 Background

As a constellation of interrelated technologies has matured in recent years, geospatial land use/land cover (LULC) information has emerged as a key input to decision-making for a host of actors, from national policymakers to urban planners to disaster relief organizations. Strikingly simple at its core—showing what lies where and when—LULC information has a wide and ever-expanding range of applications. In the urban context, and with a sufficient classification scheme, these could include:

- detecting changes in the size and distribution of green and open spaces within urban areas;
- quantifying the impact of different types of urban land use on air quality;
- identifying what types of land use have been affected by disasters; and
- analyzing how urban areas have developed historically, and monitoring in near real time how they are changing.

Additional examples of applications and related audiences can be found in Appendix A.

1.2 The Difficulties of Production

The challenges of producing LULC information, particularly in the context of urban spaces, are evidenced by its scarcity. Despite their utility, high-quality, regularly updated, publicly available LULC maps of cities are rare. Comparable maps of different cities created with a consistent methodology are even less common.

The primary constraint is well-known: until fairly recently, the only reliable method of urban LULC mapping that could be scaled to large areas—in contrast to precise but inherently local on-the-ground surveying—was manual classification of satellite (or aerial) imagery, a time- and resource-intensive process. Furthermore, this task never truly ends. As urban footprints expand and shift, as development and redevelopment churn, LULC continuously changes. This very change is one of the most valuable dimensions of LULC information, but also the most difficult to capture, as hefty resource demands are joined by the additional challenges of maintaining consistency over time.

Box 1 | What Are Land Use and Land Cover?

Land use and land cover are distinct but closely related concepts for describing what exists at a given location on the surface of the Earth. According to the National Oceanic and Atmospheric Administration (NOAA), "Land cover indicates the physical land type such as forest or open water whereas land use documents how people are using the land."[a]

A land use/land cover (LULC) map classifies areas within a set of categories. The classification system employed is dictated by the purpose of the map. The system can be binary, as with "built-up" versus "not built-up" in the Global Human Settlement Layer,[b] or an elaborate and extended taxonomy, like the U.S. Department of Agriculture (USDA) Cropland Data Layer,[c] or anything in between. Many classification systems include both land use and land cover. For example, the typology of the Cropland Data Layer includes "Water" and "Evergreen Forest"—types of land cover—as well as "Strawberries" and "Developed/Low Intensity"—types of land use. The classification system described in this paper is a mixed typology, which focuses on distinguishing between different types of built-up areas within cities.

Notes:
a. Additional information at NOAA: https://oceanservice.noaa.gov/facts/lclu.html.
b. Additional information at Global Human Settlement Layer: https://ghsl.jrc.ec.europa.eu/.
c. Additional information at U.S. Department of Agriculture: https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php.

1.3 The Case for Automation

Automating urban LULC classification offers the chance to reduce the resource and time requirements by orders of magnitude, while ensuring methodological consistency. At the same time, automation can dramatically expand the scope of the mapping. Whereas the most comprehensive publicly available, global urban LULC mapping project to date[1] covers approximately 1 percent of 200 large cities at three points in time since 1990,[2] algorithmic classification could provide 100 percent coverage of the roughly 10 percent of the Earth's land surface that is covered by settlements, and accounts for 95 percent of human population, on an ongoing basis (Asher 2019).

Whether automated LULC classification of urban areas can match or exceed human performance is interesting from a technical standpoint, but given the current scarcity of such LULC data, the salient question is whether an automated process can be accurate and reliable enough to produce information that is useful for decision-making. Furthermore, the extent to which the information distinguishes different land uses—basically, the precision of its classification system—dictates what decisions it can help inform.

1.4 Existing Techniques and Products

Some businesses have recently begun selling information related to urban LULC. DigitalGlobe, a privately owned satellite imagery provider based in Colorado, offers categorized building footprints: the location, outline, and type of individual structures. These data are sold on a per-building basis, but the company does not publish the price or its methodology.[3] The availability of footprint data from past points in time is also unclear. Orbital Insight, a geospatial intelligence company headquartered in Palo Alto, California, offers a land use classification service,[4] but similarly does not disseminate pricing or methodology information. Both LULC data sources rely on the high-resolution imagery that they also sell, making cost and coverage serious obstacles to any global system.

There have also been academic research efforts outside of the commercial sector. The World Bank, in collaboration with Oak Ridge National Laboratory, developed a system for characterizing formal and informal settlements (Graesser 2012). At the heart of its methodology are the "spatial, structural, and contextual features" of urban spaces—things like lines, edges, and shadows. The system yielded impressive results in a handful of large, global cities, but this so-called textural approach requires very-high-resolution imagery, limiting its scope for application. "Object-based" or "object segmentation" techniques, as deployed by, for example, Patlolla et al. (2013), Banzhaf and Höfer (2008), and Gamba et al. (2007) are similarly constrained by the need for very-high-resolution imagery to identify fine-grained features.

The CORINE Land Cover inventory of the Copernicus Land Monitoring Service (CLMS) meets many of the discussed criteria: it utilizes an extensive typology, covers several points in time across decades, and, though not global, is continental in scale (EEA n.d.a). The results are all publicly available as well. These features are shared by the Urban Atlas project, also by CLMS (EEA n.d.b). However, categories like "continuous urban fabric" and "discontinuous medium density urban fabric" speak primarily to land cover, not land use, thereby limiting their range of application in the urban context.

Other well-known data products in this area are similarly oriented around land cover rather than land use. The Global Land Survey from the National Aeronautics and Space Administration (NASA), U. S. Geological Survey (USGS), and the University of Maryland focuses on tree cover (USGS n.d.). The Climate Change Initiative Land Cover Project by the European Space Agency (ESA) deploys a rich typology with dozens of categories, but all built-up spaces are collapsed into a single "Urban areas" class (ESA 2015a). Open Street Map is one of the very few projects that incorporates both land cover and land use data, but quality and especially coverage are highly variable in different areas, making it unsuitable for systematic use.

1.5 The Emerging Opportunity

Several obstacles have traditionally impeded attempts to automate LULC classification of urban areas using satellite imagery. Although some imagery was publicly accessible and used to observe large-scale features like the Amazon rainforest, its spatial resolution was insufficient to capture the finer features present in urban environments. For example, the Moderate Resolution Imaging Spectroradiometer (MODIS), launched by NASA in 1999, has a maximum spectral resolution of 250 meters, and so cannot meaningfully capture even a sizeable roadway—a four-lane road is still only about 10 meters wide (NASA n.d.). Commercial imagery offers superior spatial resolution, but its high cost and low spatial coverage often introduce major challenges. Commercial systems typically acquire imagery on demand, so many areas are ignored or seldom captured; compounding this, fewer satellite visits means fewer chances to acquire cloud- and shadow-free imagery. Ground-truth data (see Box 2 "What Is Ground-Truth?"), which represents definitive information about the precise location and nature of features on the ground, has also been in short supply. The sheer volume of imagery and ground-truth data needed to inform an automated approach, along with the associated computing demands, have posed additional technical hurdles.

Box 2 | What Is Ground-Truth?

Ground-truth is the vital, definitive, geospatially explicit information about what actually exists at specific locations. Ground-truth data may be collected by humans on the ground who can directly attest to the nature of the LULC in a given place, or may be produced by human analysts who manually examine high-resolution satellite imagery (or aerial or drone photography) and ascribe particular LULC categories to precisely demarcated swathes of land. Our project utilizes ground-truth produced in the latter fashion. For more information, please see sections 2.3.1 "Atlas of Urban Expansion" and 2.4.1. "Constructing Training Data."

Its name notwithstanding, "ground-truth" is rarely perfect. Like any other data, ground-truth may contain errors as well as flat-out inaccuracies. This paper refers to these data as "definitive" because we conduct our modeling as if they are correct—for example, when evaluating the accuracy of model predictions. We are not modeling reality; we are modeling data we believe to be representative of reality. The utility of the model extends only so far as that assumption holds true.

But this situation has changed. Imagery from Landsat[5] satellites is now in the public domain. Furthermore, the European Union's Copernicus Programme[6] provides regular global coverage and features improved spatial and spectral resolution over previous satellite programs. New machine learning techniques have matured and now underpin automation across a host of domains, from predictive internet searches to autonomous vehicles. Both driving and being driven by these advances, commercial cloud computing has exploded in scale and plummeted in price, putting huge amounts of processing power and storage within reach at modest cost.[7] Lastly, new, open sources of ground-truth have emerged, providing crucial geospatial LULC data for automation efforts to reference.

2. METHODOLOGY

2.1 Requirements

Our methodology has two basic requirements. First and foremost, we can characterize LULC only at times and places for which cloud-free satellite imagery is available. Second, we need ground-truth to create or improve classification models.

Many machine learning systems "learn" to recognize patterns by "looking at" training samples, meaning inputs and corresponding outputs. The system iterates over such examples and continually adjusts its underlying mathematical algorithm—the model, per se—in order to link inputs to expected outputs ever more closely. This is known as "supervised learning."

In our case, a typical training sample consists of satellite imagery of a given location and the actual LULC at that location—the model input and expected output, respectively (see section 2.4, "Model Creation"). Without both parts, the system cannot learn to associate various appearances with particular LULC categories—the essence of the project.

Note, however, that once trained, a model can be applied to *characterize LULC anywhere that imagery is available.*[8] This ability to create comprehensive LULC maps extending far beyond limited ground-truth coverage is key. Because of the unrivaled temporal and spatial coverage of medium-resolution imagery, thanks to missions like Sentinel and Landsat, our architecture is theoretically capable of generating truly global maps of urban areas across decades. With that longer-term possibility in mind, we set out to test the feasibility of producing LULC classification and change-detection products that maintained fidelity and consistency across a diverse set of urban landscapes while using only publicly available data and open-source algorithms.

2.2 City Selection

The primary source of ground-truth for this project was the Atlas of Urban Expansion.[9] An earlier pilot phase and subsequent testing suggested that our mapping algorithms struggled the most in the Indian subcontinent, relative to their performance elsewhere. Therefore, in order to develop the most robust and capable models, we focused on that region for the research described in this paper. The Atlas includes 17 Indian cities. After accounting for satellite coverage, prevailing atmospheric

conditions, and platform limitations, we focused on the 11 for which an adequate amount of clear imagery was available: Ahmedabad, Belgaum, Hindupur, Hyderabad, Jalna, Kanpur, Parbhani, Pune, Singrauli, Sitapur, and Vijayawada. Although these cities constituted our proving ground, the methodology itself is not contingent on geography. We can create and train models anywhere we have imagery and ground-truth, and we can apply models to generate LULC maps anywhere we have imagery. However, we can only "test"—that is, quantitatively assess—model performance in places where we have ground-truth data that allow us to compare model prediction to reality.

2.3 Data Sources

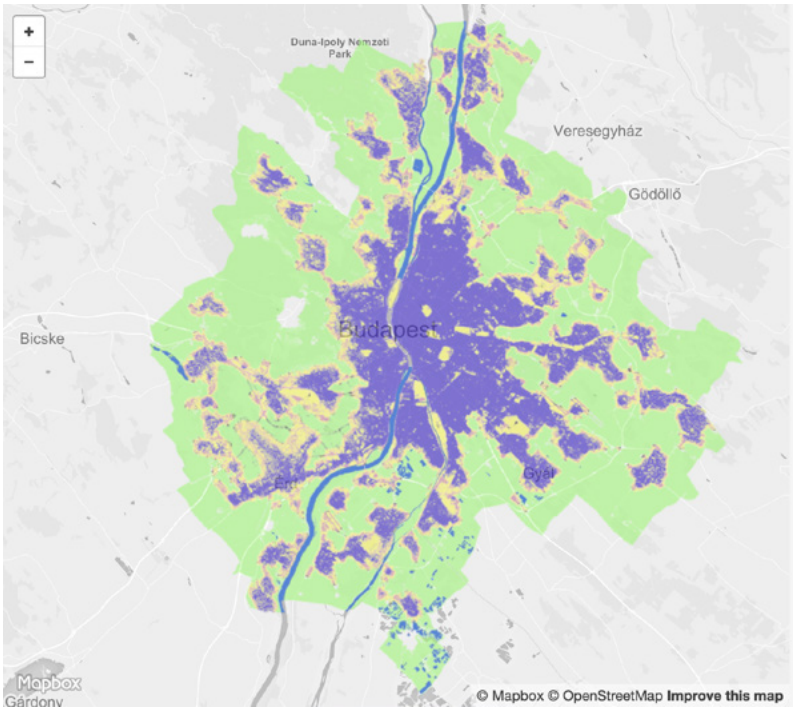
The project relied on two key sources of data: the Atlas of Urban Expansion for LULC ground-truth and the Sentinel-2 satellite constellation, accessed via the Descartes Labs platform, for satellite imagery.

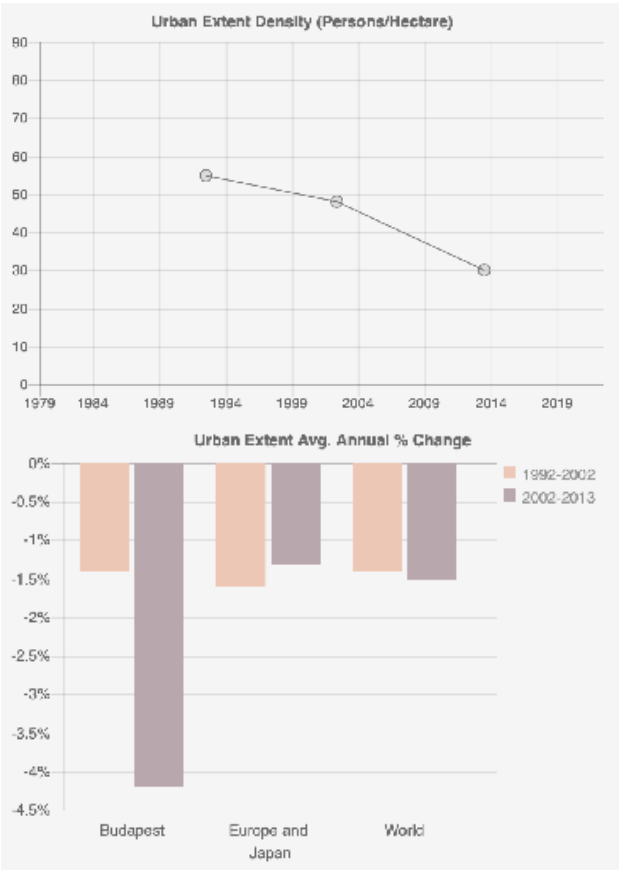
2.3.1 Atlas of Urban Expansion

The Atlas of Urban Expansion is a multiphase research effort conducted by the New York University (NYU) Urban Expansion Program in partnership with UN-Habitat and the Lincoln Institute of Land Policy. The Atlas, which analyzes a global sample of 200 cities, aims to examine how and why urban areas and their peripheries have changed over time.

The first phase of the Atlas project focused on mapping and measuring several key attributes of urban expansion at three points in time in the last 25 years, while the second evaluated the characteristics of recent and older urban layouts. This entailed manual digitization and classification of medium-resolution Landsat imagery, as well as high-resolution satellite mosaics accessed via Bing and Google Earth. The resulting data enabled longitudinal and cross-sectional analyses, yielding unprecedented quantitative characterization and comparison of the urban form of the 200 sample cities (see Figure 1).

Figure 1 | **Budapest Study Area and Comparative Statistics as Seen on the Atlas of Urban Expansion**





Note: The full city profile for Budapest can be viewed on the Atlas of Urban Expansion website: http://atlasofurbanexpansion.org/cities/view/Budapest.
Source: Angel et al. 2016a.

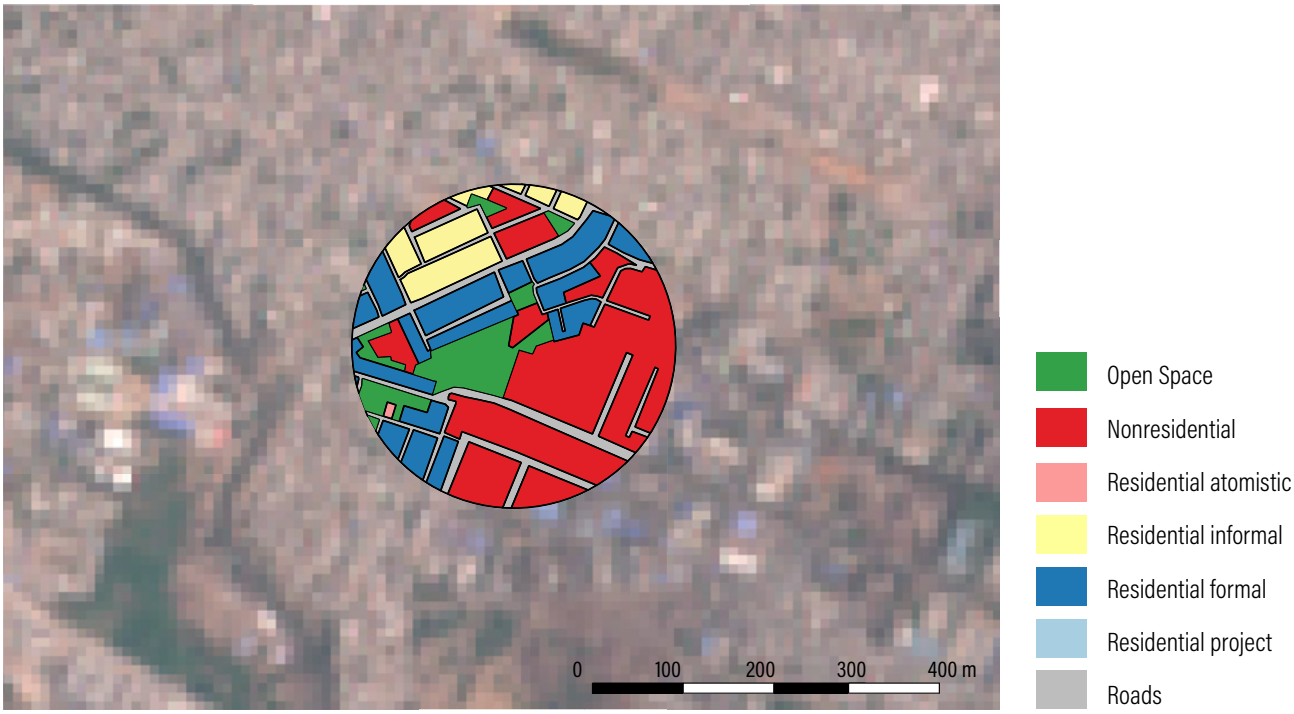
To our knowledge, the Atlas represents the most detailed, well-documented, and scientifically rigorous effort to systematically classify land use in cities on a global scale. In particular, the design and clarity of the underlying methodology make the Atlas especially well suited for our project, as do the heterogeneity and geographic diversity of its sampling of large cities. A comparable product like the European Environment Agency (EEA) Urban Atlas similarly contains a wealth of high-quality LULC data, but is less suitable for our purposes, since the dataset is continental rather than global, and employs a much deeper set of categories, which simply may not be distinguishable in medium-resolution imagery (EEA 2017). Evaluation of the relative effectiveness of other sources of ground-truth as input for training data fell outside the scope of our research.

2.3.2 Descartes Labs

Descartes Labs[10] is a small U.S.-based company, spun out of Los Alamos National Laboratory, that develops commercial, cloud-based algorithms and software to monitor the world's resources using overhead imagery and other complex datasets. Descartes Labs partnered with the World Resources Institute (WRI) to demonstrate its technology on a range of real-world applications with high potential impact for WRI's mission.

Descartes Labs has implemented a cloud-based software platform for daily ingestion, cleaning, and calibration of public-domain satellite imagery, including all data from Landsat, Sentinel-1, and Sentinel-2.[11] This platform for turning satellite observations into analysis-ready data is highly scalable, having reprocessed the entire Landsat and NASA MODIS archives—over one petabyte of uncompressed imagery—in under 16 hours (Warren et al. 2015). The registration and calibration routines employed by the platform are the official open source algorithms published by the Landsat and Sentinel scientific communities. The Descartes platform is designed specifically for automated, large-scale retrieval and analysis of satelite and aerial imagery, which our project executes via private computer nodes within Google Cloud Platform. Our project has used that platform primarily to access 10-meter resolution Sentinel-2 multispectral imagery, but has also experimented with 30-meter Landsat multispectral imagery, 1-meter aerial

Figure 2 | **Ground-Truth for an Atlas of Urban Expansion "Locale"**



Note: Atlas of Urban Expansion data (laid over a Sentinel-2 satellite image) for a single "locale" in Hyderabad, India. Polygons correspond to manually identified contiguous areas of a single LULC type. Each locale has a radius of approximately 178 meters. (This locale includes all categories but "Residential project.")
Source: Land use polygons from Angel et al. 2016b. Graphic by authors.

photography, and 20-meter Sentinel-1 synthetic aperture radar data.

2.4 Model Creation

Machine learning uses training data to establish relationships; in this case, between how an area appears in satellite imagery and its LULC classification. These patterns are captured within models—in essence, mathematical equations—that take as input the pixel values from satellite imagery at a given location, and return as output a LULC category (e.g., "residential"). We used supervised learning methods, in which a model "learns" from training samples, each of which links an example input to the corresponding, actual output. The process of model creation thus begins with training data construction.

2.4.1 Constructing training data

GROUND-TRUTH

The Atlas of Urban Expansion project makes its data freely available[12] via a large number of geospatial files for each studied city. These represent datasets characterizing various dimensions of the study area: the extent of the urban space at different points in time, the arterial road network, and so on.

Two such datasets are essential to our work. One outlines the study samples, or "locales": 10-hectare circles quasi-randomly selected from throughout each studied city. Collectively, these circles cover roughly 1 percent[13] of an urban area.[14] This is substantial: the dataset for a large city like Mumbai includes hundreds of these locales, constituting 20–30 square kilometers. We leveraged every square meter of these data, either as training or validation data. The second dataset consists of polygons contained within these locales. As shown in Figure 2, these polygons employ the Atlas's typology to classify LULC across the entirety of each locale. These polygons were manually created by a team of satelite analysis experts contracted by the Atlas of Urban Expansion project and trained with detailed guidelines to foster rigor and consistency. The polygons represent manual interpretation of high-resolution satellite imagery dating from 2013 to 2015, depending on the city. Each area is characterized once and only once—there is no temporal component to these LULC data.

The typology expressed by the Atlas LULC data is an important aspect of this project. Discriminating between types of LULC within urban spaces is a critical, distinguishing feature of our work. Precise distinctions between types of land use opens up a range of potential applications for urban planning and research that would be unavailable with a product limited to land cover. A detailed breakdown of the classification system can be found in Atlas materials; here a brief overview will suffice.

- Open Space—any lot of pervious land cover, including farmland (though not, say, a grass lawn in front of a house—the entire plot would be considered residential).
- Nonresidential—commercial and industrial usage, like warehouses, stores, factories, and air- or seaports.
- Residential—The Atlas has a particular focus on housing, hence the residential subcategories, whose precise definitions are complex, but which seek to encapsulate various levels of apparent planning and infrastructure.
- Roads—covers all roadways, paved or otherwise.

These polygons are the foundation of our ground-truth: they explicitly ascribe LULC to precisely demarcated areas. However, preparing them for machine learning ingestion required a significant amount of processing.[15] Using GIS software, we consolidated hundreds of geospatial files, manually reconciled overlapping or malformed geometries, and added explicit polygons for roadways (which originally were signified only by the gaps between LULC polygons). The output of this process was a single geospatial file per city. Its contents comprised a comprehensive, spatially precise classification of LULC within all locales for that city—meaning that every location within every locale was definitively ascribed a LULC category. The final step, in which these vectors were converted into raster data, is described in the subsection below, "Combining imagery and ground-truth."

SATELLITE IMAGERY

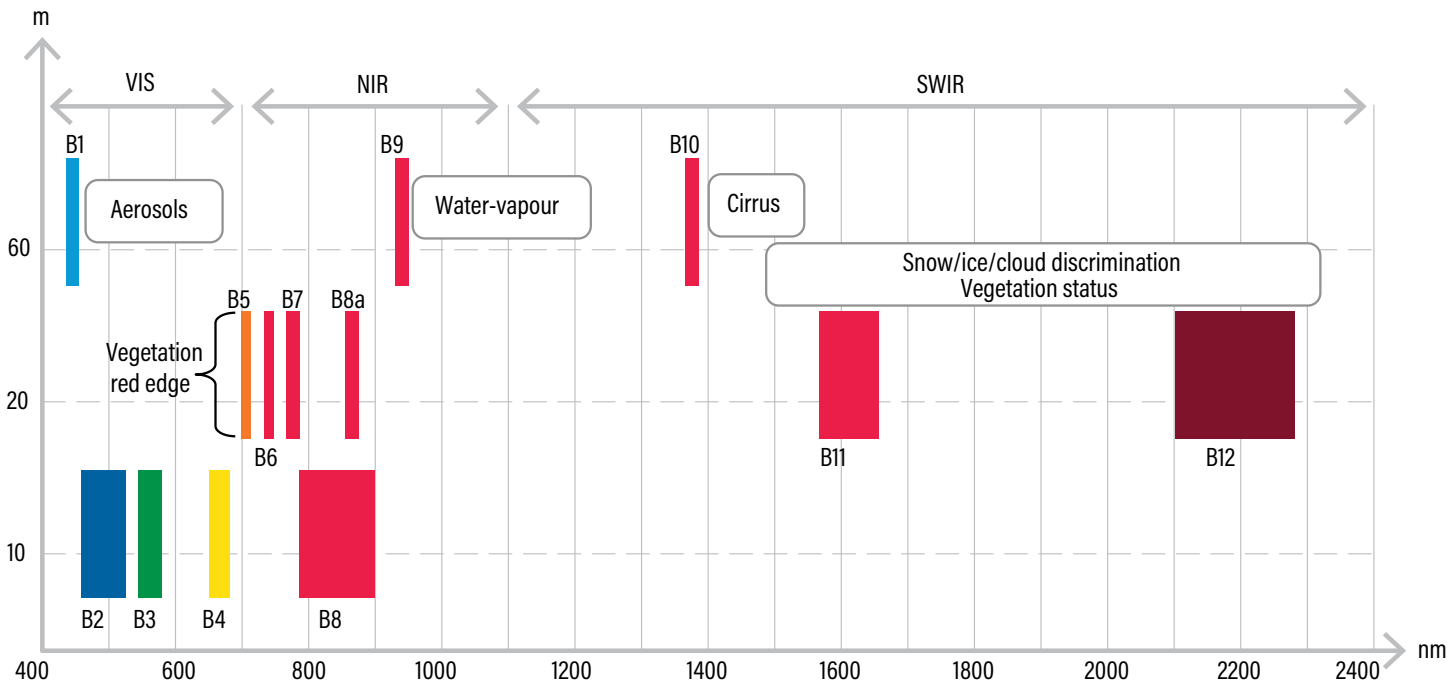
Imagery is the other essential input to our machine learning process. Unsurprisingly, the source of imagery strongly impacts the resulting model. Different satellites carry different instruments, which in turn have very different imaging capabilities, from spatial resolution to spectral band availability. The resolution and bands effectively dictate what the classification algorithm can and cannot "see," and thus what it can and cannot possibly characterize. Table 1 summarizes these differences, while Figure 3 gives a detailed breakdown of the imaging capabilities of Sentinel-2, our chosen source of satellite imagery.[16] Figure 4 provides an example of the

Table 1 | **Comparison of Satellites' Imaging Capabilities**

SATELLITE	LANDSAT	SENTINEL-1	SENTINEL-2	NATIONAL AGRICULTURE IMAGERY PROGRAM (NAIP)
Spatial Resolution	15–30m/pixel	20m/pixel	10–20m/pixel	1m/pixel
Available Bands	Red, green, blue, near infrared, shortwave infrared, cirrus	Dual polarization radar	Red, green, blue, near infrared, shortwave infrared	Red, green, blue (near infrared in some areas)
Revisit Time	16 days	3–6 days	5–10 days	1–3 years
Years Available	1972–now	2014–now	2015–now	2003, 2008–2018

Note: Resolution is band-dependent, with the various imaging devices able to resolve different spectral bands at different spatial resolutions (for an example, see Figure 3). The listed values represent the resolutions of the bands most relevant for our project, which typically excludes the coarsest bands. Also note that "Landsat" in fact denotes a series of satellite constellations. Instrumentation and capabilities have changed substantially over nearly five decades.
Sources: Metadata for Landsat 8 from USGS 2019. Metadata for Sentinel-1 and Sentinel-2 from ESA 2015b and 2015c. Metadata for NAIP from USDA n.d.

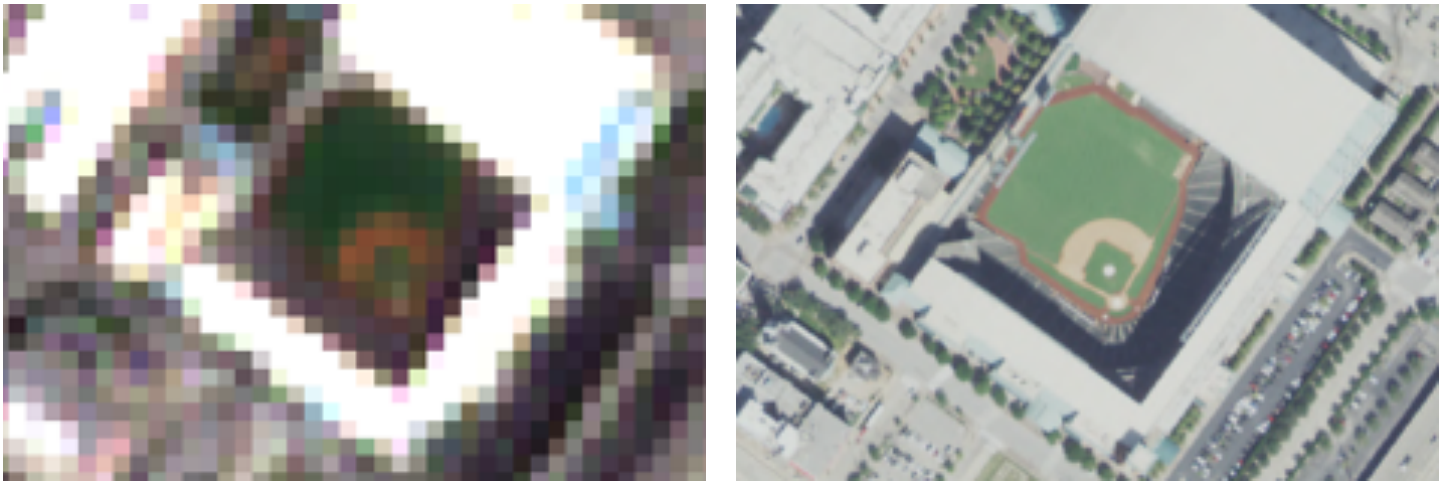
Figure 3 | **Imaging Capabilities of Sentinel-2**



Spatial resolution versus wavelength: Sentinel-2's span of 13 spectral bands, from the visible and the near-infrared to the shortwave infrared at different spatial resolutions ranging from 10 to 60 m on the ground, takes land monitoring to an unprecedented level.

Source: ESA 2017.

Figure 4 | **Comparison of Sentinel-2 and NAIP Imagery**



Note: On the left, Sentinel-2 10m/pixel imagery of the ballpark in downtown Houston; on the right, the same ballpark as captured by NAIP at 1m/pixel spatial resolution.
Source: Sentinel-2 imagery from the European Space Agency. NAIP imagery from U.S. Department of Agriculture. Imagery accessed using Descartes Labs Viewer.

difference in spatial resolution between imagery sources, including the nonsatellite National Agriculture Imagery Program (NAIP), which conducts aerial photography.

Spatial resolution is particularly significant for our project, which focused on fine urban features like roads. Our images are rasters: essentially, grids of pixels. The spatial resolution of the imagery determines the size of the grid elements, and thus the area of the Earth's surface captured by each pixel. Higher resolution means smaller pixels, resulting in a finer grid, whereas lower resolution means larger pixels and a coarser grid. When we refer to a "location" or ascribe a LULC to it, we refer to a pixel, and the entire area it covers. Although subpixel classification is a long-standing topic of research, for the purposes of this paper, the spatial resolution of our imagery constitutes a hard ceiling on the precision of our analysis and LULC maps we generate.[17]

Additionally, the temporal resolution, or revisit time, and the historical record of an instrument (the years from which imagery is available) have significant bearing on its potential uses and the timeliness of the insights that can be derived from its observations. For example, Sentinel-2's combination of high temporal resolution and short historical record makes it an excellent candidate for timely monitoring of recent and future changes, but not a viable option for longitudinal research looking back more than a few years.

Nearly all image preprocessing was performed by the Descartes Labs platform. We used native platform functionality to exclude heavily clouded satellite captures, and to select and combine multiple overlapping, contemporaneous captures into unified mosaic images that fully covered the study areas.[18] Out of the enormous pool of available imagery, the specific images used for training were manually selected by our team. Curation was a matter of direct inspection of each image, looking in particular for obscuring clouds (or their shadows) over areas containing ground-truth and for washed-out coloration indicative of atmospheric haze. Other potential imagery flaws, such as geometric distortion, were not systematically evaluated.

Ultimately, we utilized Sentinel-2 imagery (the Level-1C product, representing top-of-atmosphere reflectance), with individual images manually selected. We did not "composite" or otherwise combine images, with the exception of mosaicking multiple images when necessary to get a picture of the full study area. In these cases, the mosaic constituents were always contemporaneous, typically captured on the same day.

COMBINING IMAGERY AND GROUND-TRUTH

Our machine learning algorithm characterizes the LULC at a location—that is, for a single pixel—by examining the imagery at and around that point—an area we refer to as the "look window." This local imagery is expressed via a

numeric array, whose values correspond to the brightness in each spectral band for each pixel in the look window.

To align the ground-truth with the satellite imagery, we transformed the LULC polygons from vectors into rasters, with pixel size matching the spatial resolution of the underlying imagery, as illustrated in Figure 5. If polygons representing multiple LULC types fell within a single raster cell, the pixel was assigned the category that covered the largest fraction. This rasterized version of the ground-truth is intended to represent the best possible LULC classification at each location/pixel for the resolution available.

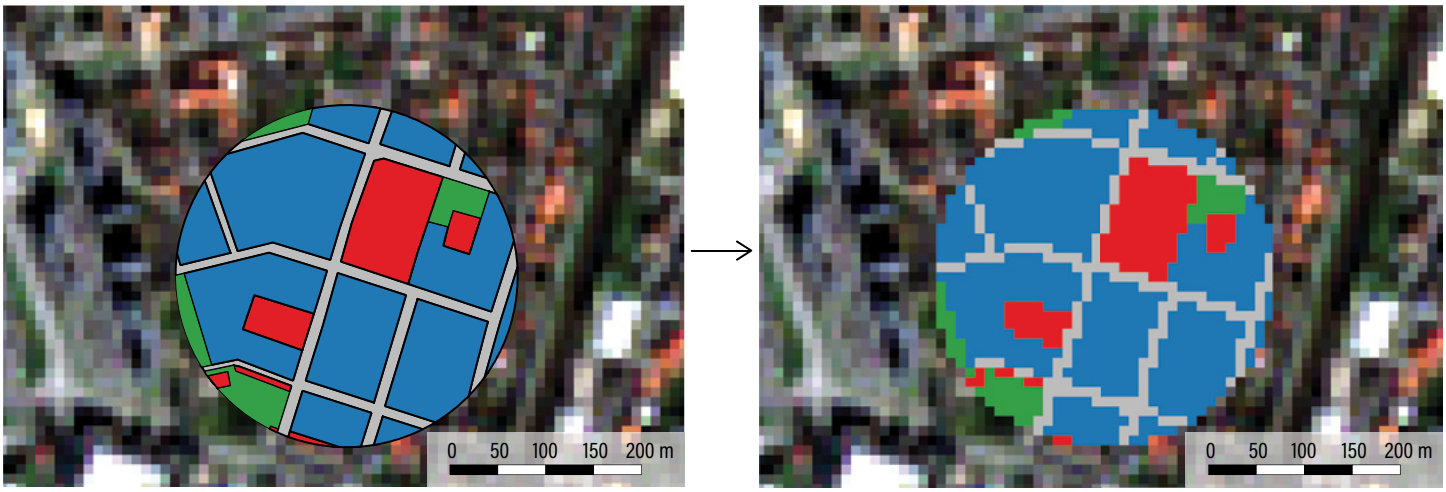
When the multi-band, multi-pixel numeric array representing the look window around a pixel is joined with the known LULC at that location, that combination constitutes one training sample. When repeated for every pixel with a known LULC classification, we have a set of inputs and the corresponding desired outputs: our training data.

OUR TRAINING DATA CONFIGURATION

Over the course of our research, we explored a wide variety of training data configurations. In the pilot phase of the project, we trained models using imagery from from Sentinel-2, Sentinel-1, and several generations of Landsat satellites, as well as composite images incorporating multiple satellites; when modeling U.S. cities, we also tried NAIP imagery. In each case, we tested a range of spatial resolutions—upsampling medium-resolution Landsat and Sentinel data to 15, 10, and 5 meters and downsampling high-resolution NAIP images to 2 and 5 meters—and look windows, expanding or reducing the physical area around a given pixel that was available to the model. Most sets of training data included visible spectral bands and various segments of the infrared spectrum. Beyond these directly observed inputs, we also tested a variety of spectral indices[21] and other derived inputs. In this exploratory phase we did not systematically investigate nonimagery inputs.

All results reported in this document were produced using a single, consistent configuration, chosen for performance in previous trials. Training data were built from Sentinel-2 imagery—as mentioned, Sentinel-2 offers the best available combination of revisit rate, spatial resolution, geographical coverage, and public accessibility—uniformly upsampled to 5 meters per pixel, in hopes of utilizing as much information as possible from our high-precision ground-truth.[22] Training sample inputs comprised the following bands: blue, green, red, near infrared (NIR), shortwave infrared 1 (SWIR1), and shortwave infrared 2 (SWIR2).[23] As an additional pseudoband, we included the normalized difference vegetation index (NDVI),[24] derived directly from the imagery.[25] Ground-truth vector files were transformed into geospatial rasters precisely coterminous with the satellite imagery rasters using the Geospatial Data Abstraction Library (GDAL).[26] From each image used for training, for every location in the ground-truth raster

Figure 5 | **Rasterization of Atlas of Urban Expansion Ground-Truth**



Note: Atlas of Urban Expansion LULC polygons atop Sentinel-2 image (left), along with LULC polygons rasterized at the spatial resolution of the underlying imagery (right), allowing each image pixel within the locale to correspond to a single LULC category. Together each pixel and corresponding ground-truth LULC classification constitute one training sample.
Sources: Underlying imagery from Sentinel-2 accessed via the Descartes Labs platform; LULC polygons from the Atlas of Urban Expansion; graphic by authors.

with a known LULC, we extracted a 17x17 square of pixels centered around that location, resulting in a look window 85 meters to a side.[27]

2.4.2 Training models via machine learning

"Machine learning" frequently refers to teaching computers to perform tasks by providing examples, rather than through explicit programming. In our project, this was accomplished by iterating over the training samples constructed as described above. Importantly, the set of samples used for training a model may come from one image of one city, many images of one city, or many images of many cities. For each training sample—each pixel with a known LULC classification—the input values are fed into the equation that constitutes the model. The output of that equation is the model's prediction for the LULC classification of that sample. This prediction is compared to the ground-truth, and the equation is slightly adjusted, in accordance with the accuracy of the prediction. The process then repeats.

The structure of the model and the way it modifies itself (i.e., the way it "learns") depend on the machine learning context: the nature of the task, the amount and quality of available training data, and so on. Furthermore, the final result of the machine learning process—a trained model—is impacted by a host of numeric parameters, which determine things like how quickly it adjusts its underlying equation in response to an incorrect prediction and whether, as well as how much, to emphasize certain types of training samples.

OUR MACHINE LEARNING CONFIGURATION

As with training data construction, we tested a wide range of machine learning structures and parameters over the course of our research. However, all results presented in this document were generated by models with an identical configuration, meaning they shared the same initial mathematical structure and the same training regime. From city to city and model to model, only the training samples changed—which of course resulted in distinct final models.

We created city-specific models,[28] each trained using samples drawn from a single city. The training dataset for each was constructed from several images[29] of the city, all captured within a single season[30] of a single year[31] —the post-monsoon period of 2017—and each of which covered the entire study area. For each image, every available pixel of rasterized ground-truth was paired with the corresponding pixel of imagery, meaning that the final training dataset contained a number of samples for each ground-truth pixel equal to the number of images used. This ran between four and eight, depending on the size of the city, and was ultimately constrained by the total number of samples that could fit in the computer's memory simultaneously. Prior to machine learning ingestion, the training data were preprocessed[32] to facilitate faster training. In every training, the relevant set of samples was divided[33] into two tranches: 70 percent were used for actual training, while 30 percent were withheld as validation data.

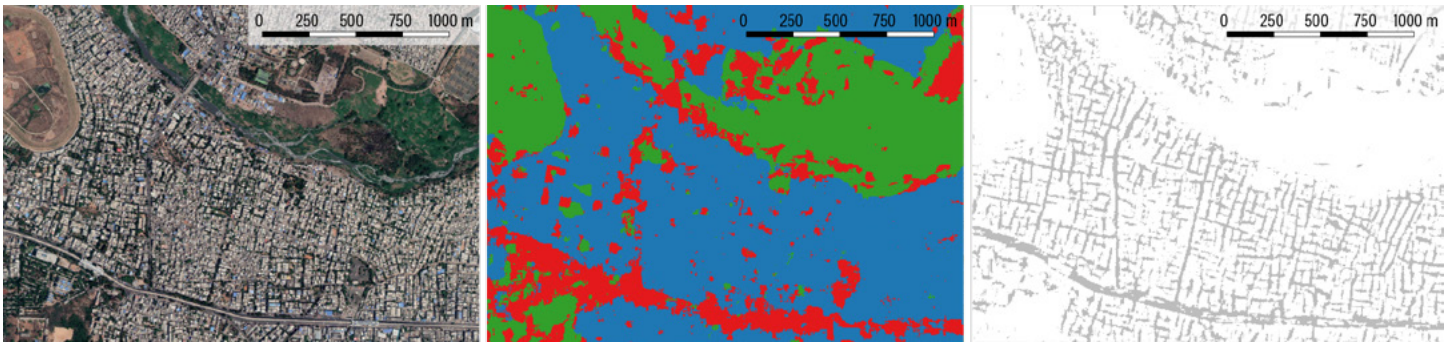
All model development was conducted using data from just two cities, one large and one small: Hyderabad and Hindupur. All of the hyperparameter selection and other experimental variations described above were guided by model performance in these two places. Therefore, the validation tranches for the other nine cities better represent out-of-sample data, and effectively serve as our test data.

The models themselves were small convolutional neural networks[34] (CNNs), implemented using the Keras library[35] running on top of TensorFlow.[36] Each model was trained quickly using a high learning rate, and then underwent additional, slower training against the same data with a much lower learning rate. Training was concluded when the model stopped improving, as determined by performance in classifying the withheld validation samples. Class distributions varied greatly from city to city, though open space was typically the dominant category. We compensated for class imbalances by using a loss function that attributed greater importance to samples from minority classes.

In our final configuration, we created two models for each city. The first would characterize LULC using a three-class typology: open space, nonresidential, and residential.[37] Rather than attempting to train models to distinguish between the four classes of residential LULC in the Atlas's typology—not all of which are necessarily present in any given city—we aggregated all types of housing into a single category. The trained "3-category model" was then applied to imagery to classify every pixel within the urban extent as one of these three types of LULC.

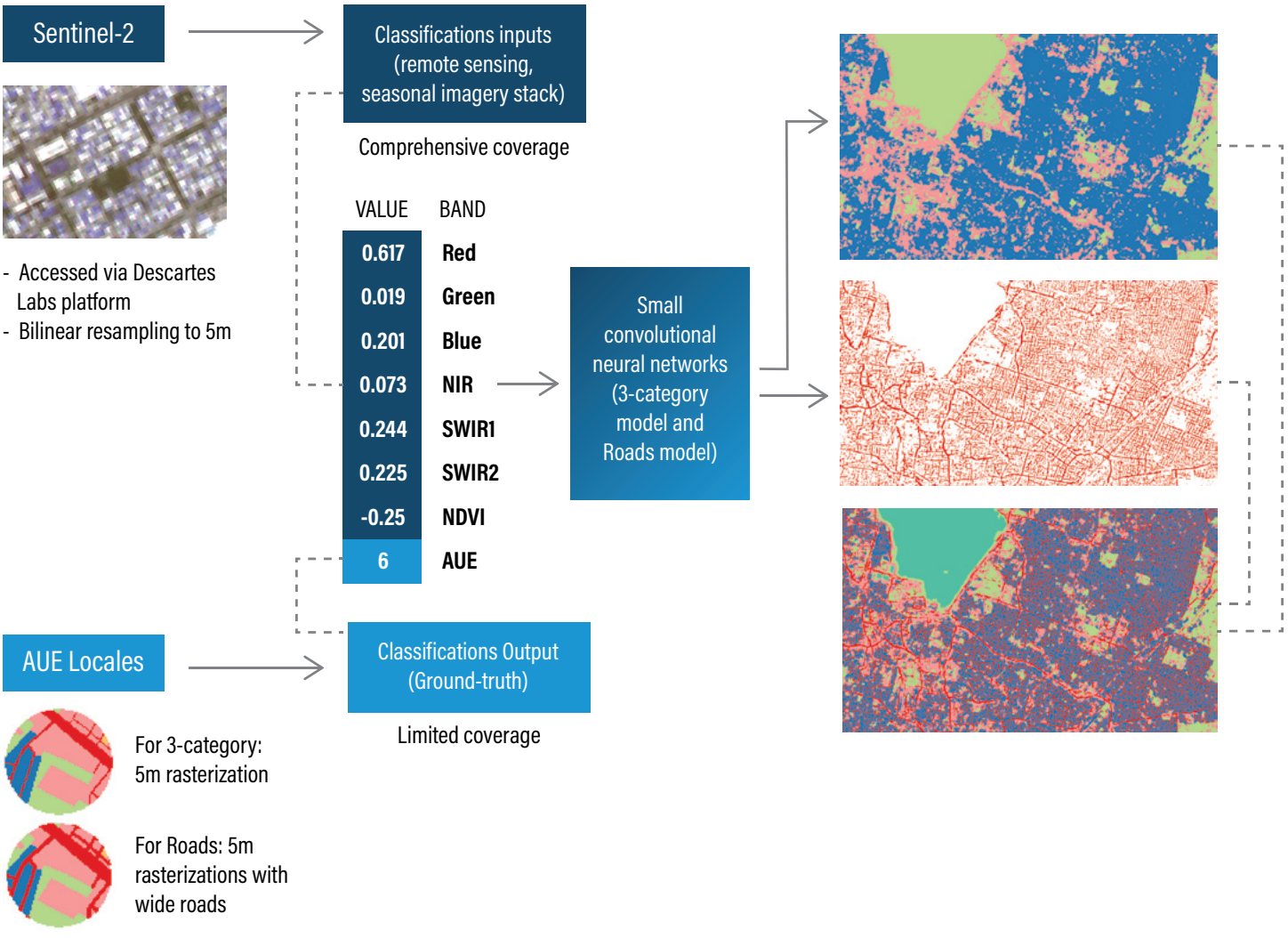
The second model was trained to make a binary determination for every pixel: road or not road. This decision was motivated by the qualitative distinction between highly linear, structured roadways and the typically amorphous, areal other types of LULC in our typology. Just as various residential LULC types were

Figure 6 | **Hyderabad Imagery and Generated Maps**



Note: A small section of Hyderabad as seen in high-resolution satellite imagery (left), a model-generated 3-category LULC map, where green, red, and blue correspond to open space, nonresidential, and residential, respectively (middle), and a model-generated map showing only roads (right).
Sources: Imagery accessed via the Descartes Labs platform; maps by authors.

Figure 7 | **Summary of Data Preparation and Modeling Workflow**



Source: Authors.

collapsed into a single category for training the other model, here we collapsed all LULC types other than road into a single, not road category. After training,[38] this "roads model" was applied to imagery to classify every pixel within the urban extent as road or not road.

Figure 6 presents an example of the output of these two models. Figure 7 summarizes our modeling workflow schematically. All code comprising that workflow is available in our Github repository.

2.4.3 Assessing model performance

As mentioned, not all training samples were used to train the models. Instead, in each city, about one-third of all pixels within the study area were randomly selected and held in reserve. That is, samples corresponding to these locations were never used for training, in any iteration of any model. Once the machine learning process was completed, a trained model was tested on this withheld set of validation data. Classifying LULC for samples it had never "seen" before revealed to what degree the model simply reflected the training data, and to what degree it captured more meaningful, broadly applicable patterns. All training samples were drawn from imagery from a single year, so validation scores for other years show model performance on completely novel imagery of locations that were not part of training. Separation of training and validation samples by pixel, rather than locale, may have somewhat blurred this division, as look windows for training samples could overlap look windows for validation samples.

The most straightforward metric for performance was a simple accuracy score: the percentage of samples in the validation data that a model correctly classified. This metric is easy to communicate, but has significant flaws when dealing with imbalanced data sets, where certain categories are much more prevalent than others. For example, if 90 percent of samples fall within a single category, then a simplistic classifier can achieve a high 90 percent accuracy by assigning every sample to that dominant category, without actually modeling the data in any meaningful way. As such, we also used the F-score[40] (or F-measure) on a category-by-category basis. More precisely, we used the F2 score, which summarizes performance while attributing twice as much significance to recall as to precision. Like accuracy, the F2 score maps performance from 0 to 1, with 1 signifying perfect performance.

For every such application of a model, we also generated a confusion matrix, as seen in Table 2, which tallies how many samples belonging to each category were classified by the model as members of each possible category. This allowed us a more nuanced view of model performance. For example, whereas the overall accuracy score for a model might simply indicate mediocre performance, its confusion matrix could reveal that the model was generally successful, but incorrectly classified most "road" samples, dragging down the overall score. This type of insight in turn helped refine the training process, whether that meant including new spectral bands to help models distinguish between particular types of LULC, consolidating multiple categories into a single grouping, or taking other corrective action.

We tested the performance of each model when applied to the following four groupings of samples:

- Training – the samples on which the model was trained (one test per city).
- Validation – the samples used for validation during training, drawn from the same images of the same city as the actual training samples (also one test per city).
- Same city – different samples drawn from images that were not used in training, but that capture the same city as the images that were part of the training data (four tests per city, one for each year in our time series).
- Other city – completely novel samples drawn from images of entirely different cities (40 tests per city: 10 for every other city, for each year in our time series).

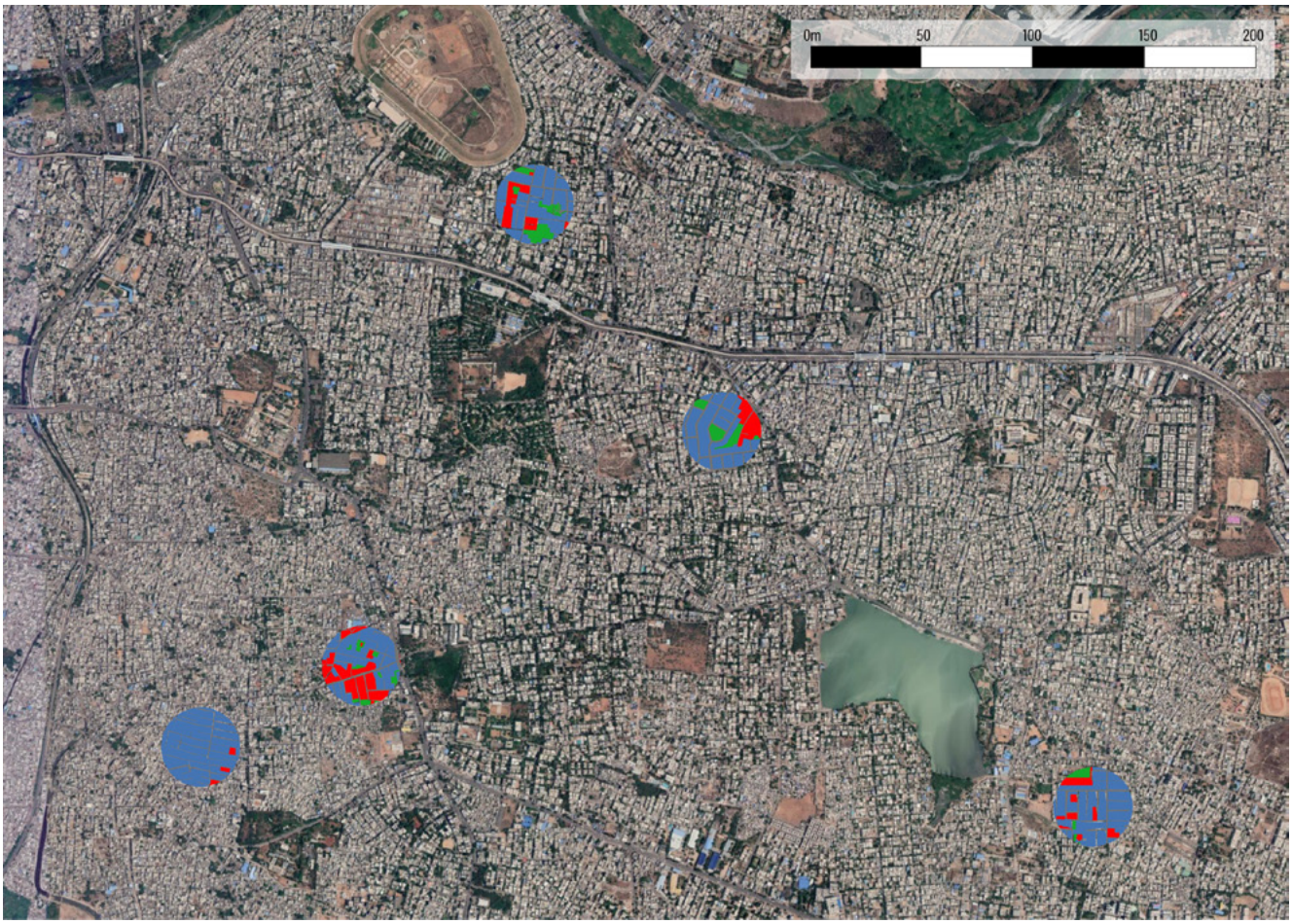
Despite the fact that during this phase of research the individual models were trained one city at a time (as described in the subsection, "Our machine learning configuration"), the eventual need for generalization and broader application was always kept in mind. Accordingly, our search for the best-performing configuration never narrowed to a city-by-city perspective. We were not trying to ascertain, independently, the best parameters to use to characterize Hyderabad, and the best for Kanpur, and so on. Instead, we focused on identifying a single machine learning configuration that yielded high-quality models in each of the Atlas of Urban Expansion cities in India.

Table 2 | Example Confusion Matrix

LULC CATEGORY		MODEL		
		OPEN SPACE	NONRESIDENTIAL	RESIDENTIAL
GROUND-TRUTH	OPEN SPACE	**21,451**	3,987	2,897
	NONRESIDENTIAL	2,403	**2,392**	2,173
	RESIDENTIAL	2,886	3,482	**18,275**

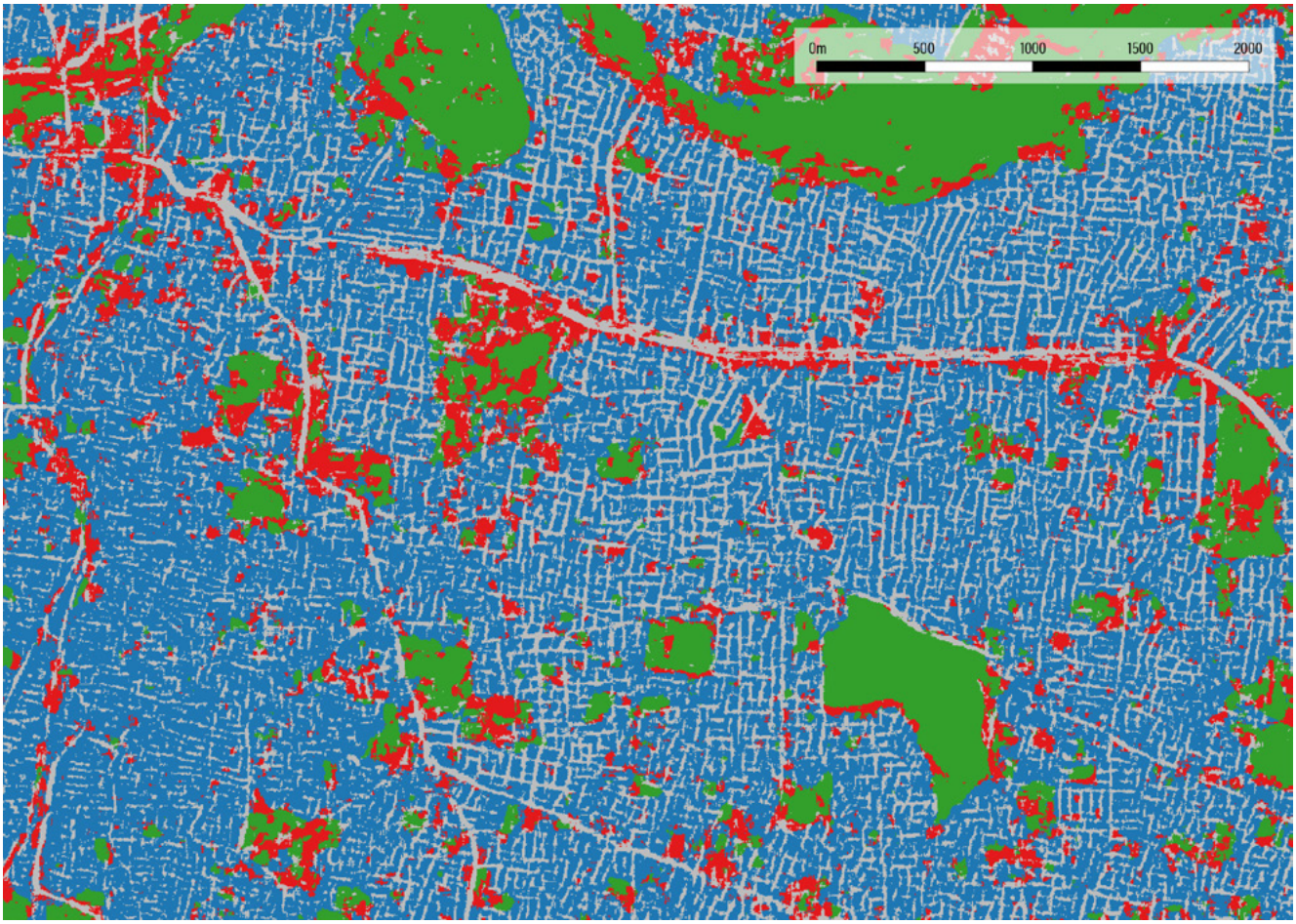
Note: The confusion matrix from application of a model to sample data, where each cell shows the number of training samples classified by the model as each type of LULC, disaggregated by the actual LULC category of each training sample. For example, there were 3,987 training samples (i.e., pixels) that the model classified as nonresidential but were in fact open space. The diagonal between the top-left and bottom-right, shown in bold, represents training samples to which the model and the ground-truth ascribed the same LULC type—correctly classified training samples. The confusion matrix of a "perfect" model would show non-zero values only along this diagonal.
Source: Authors.

Figure 8 | Detail of Hyderabad Imagery (Sentinel-2)



Note: High-resolution satellite image of part of Hyderabad, overlaid with LULC data from the Atlas.
Sources: Imagery from Google Satellite map service, accessed via QGIS; LULC polygons from Atlas of Urban Expansion; graphic by authors.

Figure 9 | **Detail of Hyderabad Imagery (Sentinel-2)**



Note: Detail of city-wide LULC map of Hyderabad (with legend), corresponding to area shown in Figure 8, which was generated by models trained using imagery and ground-truth data as seen in that figure. This graphic shows the output of the Hyderabad roads model overlaying the output of the 3-category classifier.
Source: Authors.

2.5 Model Application

As suggested by the validation process, the utility of the model lies not in its ability to learn, but in its predictive power. Accordingly, whenever a model's performance metrics suggested that it held some meaningful predictive power, we took that fully trained model and applied it to the entire city. For every point in the satellite imagery, we input the local pixel values and recorded the model's prediction, generating a comprehensive LULC map, as seen in Figures 8 and 9.

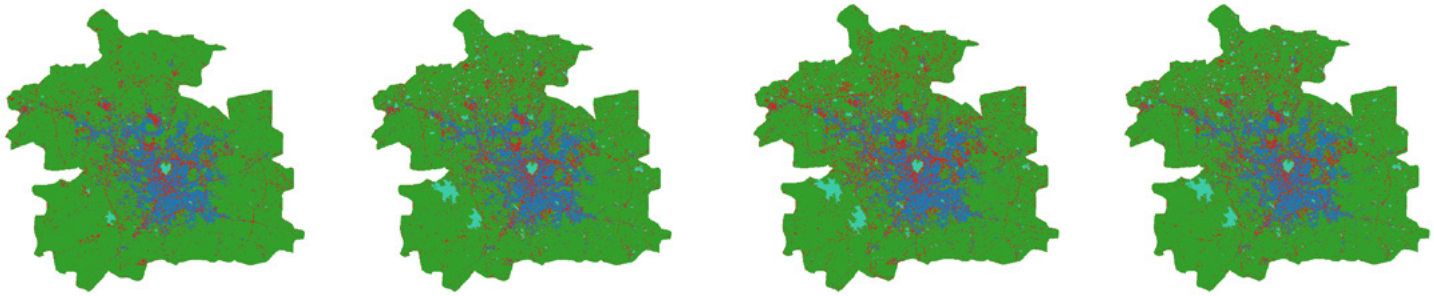
3. SAMPLE APPLICATION AND RESULTS

We created and trained two models—one for identifying only roads and the other for distinguishing between open space, nonresidential, and residential areas—for 11 Indian cities using the training data and machine learning configurations described above: 5-meter per pixel upsampled Sentinel-2 imagery; blue, green, red, near infrared, and two shortwave infrared bands; a look window 17 pixels to a side; all fed into a modest convolutional neural network.[41] The training dataset for each was constructed from four to eight Sentinel-2 images[42] of the city, all captured during a two-month span in 2017 immediately following the monsoon.

Each model was then applied to a "new"—that is, not used for training—image of the corresponding city during the post-monsoon period for every year where imagery was available. Since our modeling focused on Sentinel-2 imagery, this meant that we produced for each city a time series of annual LULC maps covering 2015, 2016, 2017, and 2018. Figure 10 shows one such map series.

Each map was complemented by statistical performance metrics, which compared the model's predictions for LULC with the human-collected ground-truth for that city. This admittedly entailed comparing predicted LULC in 2015, 2016, 2017, and 2018 with ground-truth as of 2015 or earlier, but time series ground-truth data

Figure 10 | **LULC Maps for Hyderabad in 2015, 2016, 2017, and 2018**



Note: The images here are, for each year, the 3-category result overlaid with the roads result, both overlaid with a water mask.
Source: Authors.

Table 3 | **Model Performance (F2 Score) by City and Year**

CITY	YEAR[a]	F2 SCORE			
		3-CATEGORY MODEL			ROADS MODEL
		OPEN SPACE (%)	NONRESIDENTIAL (%)	RESIDENTIAL (%)	ROAD (%)
Ahmedabad	2015	88.5	84.6	94.9	72.7
	2016	89.7	88.1	95.8	73.4
	2017	95.1	95.7	97.4	79.4
	2018	87.9	88	95.2	70.3
Belgaum[b]	2016	89.7	87.9	95.3	72.6
	2017	94	94.5	96.8	80.7
	2018	89.9	88.6	95.6	76

[a] In this table of results, most cities exhibit a discernable pattern of higher scores in 2017 than in other years. This is presumably an artifact of the model training datasets, which were created using imagery from that year only. All entries in the table represent the application of models to out-of-sample, previously "unseen" imagery; nevertheless, the out-of-sample 2017 imagery was apparently more similar to the training data than the out-of-sample imagery from other years, resulting in slightly higher model success. [b] Cloud-free Sentinel-2 imagery of Belgaum in 2015 was not available.

Hindupur	2015	82.7	61.2	82.4	32.9
	2016	83.7	71.8	85.9	52.9
	2017	93.8	94.6	95.2	81
	2018	88	79.7	88.5	61.6
Hyderabad	2015	79.8	74.5	89.9	53.1
	2016	88.2	87.2	94.2	55.5
	2017	94.4	94.8	96.8	69.9
	2018	86.7	84.7	93.3	56.2
Jaipur	2015	83.8	80.3	89.8	58.8
	2016	88.5	87.2	94	73
	2017	94.3	94.7	96.7	83.9
	2018	86.1	87.4	92.8	69.3
Kanpur	2015	82.2	69.1	88.9	55.7
	2016	87.4	82.4	92.6	67.9
	2017	89.3	88	91.9	65.1
	2018	87.8	83.4	91.4	70.4
Malegaon	2015	86.1	75.6	83.6	58
	2016	82.6	75.3	87.9	63
	2017	90.3	88.8	94.3	75.3
	2018	86.8	82.8	87.8	61.8
Parbhani	2015	82.6	59.7	88	48.1
	2016	86.4	80.4	92.2	65.6
	2017	93.2	93.9	96.8	73
	2018	86.6	78.3	92.6	53.7
Pune	2015	86.9	79.5	94.6	72.2
	2016	86.4	83.4	94.5	70.9
	2017	91.9	91.7	96.4	79.9
	2018	90.2	86.3	95.2	74.5
Singrauli	2015	74.6	71.2	79	45.4
	2016	82	84.6	87.7	67.4
	2017	82.9	79.8	86.1	71.4
	2018	79.0	78.3	84.1	61.9
Sitapur	2015	82.9	64.4	79.8	40.7
	2016	86.0	77.3	90.7	56.4
	2017	88.2	77.9	89.2	61.3
	2018	84.4	80.8	91.3	60.4

Note: F2 scores of model application to validation data from each image. Validation samples were drawn from locations not used in training. Pixels from these "unseen" locations were ingested by the model, and its predictions were then compared to the LULC classifications asserted by the ground-truth.
Source: Authors.

simply were not available. Table 3 presents the primary performance metric, the F2 score, for the validation data for every map produced for each of the cities.

Again, note that within the entries for a given city, each row represents an application of the same model to imagery from the corresponding year—with 3-category and roads being characterized by separate models. For example, the "3-category model" scores in the rows for Ahmedabad represent the application of a single model trained on 2017 images to different satellite images of the same city captured in each year from 2015 to 2018.

Table 4 lists the F2 scores from the application of the Hyderabad models (3-category and roads) to each of the four groupings. As expected, the models perform best against the training data, and scores drop off as the target samples progressively become less similar to those training samples. The models were able to predict LULC within Hyderabad—irrespective of image, year, or part of the city—with scores of 80 percent for open space, 74 percent for nonresidential, 90 percent for residential, and 55 percent for roads. For other cities, on which the models were not trained, the lowest scores fall to 57 percent, 38 percent, 49 percent, and 30 percent, respectively.

Table 4 | **Model Performance (F2 Score) across Groupings of Samples**

APPLICATION	F2 SCORE			
	3-CATEGORY MODEL			ROADS MODEL
	OPEN SPACE (%)	NONRESIDENTIAL (%)	RESIDENTIAL (%)	ROAD (%)
Training	96.2	96.5	97.7	75.0
Validation	94.3	94.8	96.8	69.9
Same city	80–95	74–95	90–97	55–73
Other city	57–69	38–51	49–79	30–41

Note: F2 scores from applications of the Hyderabad models to four groupings of sets of samples.
Source: Authors.

4. FURTHER RESEARCH, DEVELOPMENT, AND NEXT STEPS

Both the strengths and shortcomings of the system have suggested several areas for future exploration or methodological refinement. A few priority items deemed most likely to improve performance are discussed below.

4.1 Image Brightness

An issue that largely stymied the application of models beyond their training contexts was the fluctuation in brightness from one image to the next. Because our imagery represented so-called "top-of-atmosphere reflectance," the intervening space between city and satellite impacted the images in a variety of ways. An array of factors, from time of day to the amount of water vapor in the air, guarantees substantial variability across different images of a given area, as well as between different locations. Because the inputs to the models are pixel values, the output classifications are highly sensitive to brightness anomalies and inconsistencies.

Since we generated the maps and results presented in this document, the Descartes platform has begun offering a more processed imagery product representing so-called surface reflectance, meaning it approximates the appearance of the Earth at its surface, rather than from the top of the atmosphere. This processed imagery should be much more consistent from scene to scene, so may strongly ameliorate or even resolve the highly problematic issue of image brightness variability.

4.2 Additional inputs

While evaluating the performance of a system that uses imagery to classify LULC was a critical first step, we plan to test additional input features in subsequent phases of this work. We have not yet fully leveraged all of the information contained within Sentinel-2 imagery. Several bands were initially discarded due to low spatial resolution (see Figure 3), but subsequent testing has suggested that they might in fact have high discriminatory value. Derived pseudobands, such as spectral indices, may be revisited repeatedly as we continue to explore more sophisticated

model structures. Imagery capture metadata like local time, capture angle, and sun position may likewise prove useful, particularly as we expand our model application to new geographies and LULC signatures. Nonimagery remote sensing datasets such as digital elevation models are also promising. Finally, projects such as OpenStreetMap offer a plethora of potentially relevant datasets, from road network geometries to building footprints.

4.3 Multi-city training datasets

Our models were effective at characterizing LULC in the cities on which they were trained. However, performance declined precipitously when those models were applied to other cities. This is a particularly salient issue given the overarching goal of the research agenda: to characterize LULC in every large city on an ongoing basis. Given the scarcity of LULC data, most of the world will always be "out of sample," so model performance beyond where it was trained is critical.

As described in section 3, "Sample Application and Results," characterization accuracy decreased as sample novelty increased: the more dissimilar the target data are to the training data, the worse the model performs. The natural response, then, is to reduce that dissimilarity. Since the target data—the places we want to characterize—are fixed, the obvious approach is to expand the training data. A wider variety of LULC signatures within the training data will shrink the gap between target and training samples and reduce the likelihood of the model encountering completely novel situations. Just as we incorporated multiple images of a single city to deal with variance in image brightness, we can incorporate samples from multiple cities to deal with variance in LULC signatures.

4.4 Multiple models

However, we anticipate limits to the approach of feeding a model ever-wider ranges of training data. As we seek to characterize urban spaces across countries and then continents, the diversity of urban forms and their presentation in satellite imagery may simply grow too great to be captured by a single model. At that point, we will likely transition from building one global model to building a set of domain-specific models, each of which can then "specialize" in characterizing certain types of urban areas. The essential challenges here will be two. One, how to define and separate those domains: by region, climatic zone, population density, latitude, or some combination of these and other factors, perhaps using a statistical clustering method. Two, given a group of models and a target area to characterize, which model should be used? Developing robust methods of model selection will be especially important, because for the vast majority of the world, no scoring information will be available to provide direct guidance.

4.5 Machine learning model

The results presented in this document were derived from essentially the same type of model. However, we continue to experiment with different model structures, components, learning conditions, and overall machine learning architecture. Priorities for exploration or tweaking include: additional and modified CNN layers; different learning regimes, particularly rate of introduction to new training samples; identifying the most effective groupings of cities for training multi-city models; and much larger, more diverse training datasets in general. More fundamentally, we are also investigating an image segmentation approach, in contrast to the pixel-by-pixel architecture that we have employed to date, and simultaneously testing its application to very-high-resolution imagery.

4.6 Classification stability and change detection

Even using the best possible models, no two LULC classification maps will be perfectly alike. If nothing else, minor fluctuations from image to image, whether due to precipitation or atmospheric effects or other types of noise, will result in discrepancies between the maps generated from those images. However, we want to have confidence that discrepancies from map to map are not artifacts of image noise. When that is true, the maps become useful in an entirely different way: they constitute a LULC change-detection system, where changes from map to map signify actual changes on the ground. This in turn offers a whole new range of possible applications.

Thus we need to increase not just classification accuracy, but classification stability. Even as we work to improve the models themselves, we are also exploring post-processing techniques for enhancing the raw model outputs. One promising approach toward that end is combining multiple maps into a single, composite output. By generating maps from each in a series of images, we end up with a corresponding series of "votes" on the appropriate LULC classification of each pixel. Selecting the "most popular" LULC classification for each pixel yields its own map. Early testing indicates that this composite map is superior to its constituents, both in terms of subjective quality upon inspection and in terms of quantitative performance metrics.

APPENDIX A

Table A1 | **Nonexhaustive List of Example Uses and Audiences for Algorithm-Generated Global Urban Land Use Maps**

Topic	Use Case	Audience
Open space	Identify green and open spaces within urban areas and how they are changing, in proportion to other land use types	■ The Sustainable Development Goals community can monitor progress toward targets ■ Subnational governments can identify specific areas and neighborhoods where loss of open space is happening rapidly ■ Academics and nongovernmental organizations (NGOs) can identify regional patterns and investigate potential overarching causes and impacts
Open space	Identify changes in agricultural lands within urban areas and in the urban periphery	■ Subnational and national governments can see, in near real time, how changes in agricultural lands are affecting food prices ■ NGO and private sector actors can identify opportunities for improved markets
Air quality	Monitor the impact of different types of urban land use on air quality	■ Academics and NGOs can study the economic and health impacts of different urban development patterns and associated land uses ■ Subnational governments and urban planners can identify where within an urban area pollution is generated and the greenhouse gas emissions associated with different types of land use to inform future development ■ Governments can identify specific neighborhoods with poor air quality and prioritize public health efforts
Service provision	Evaluate urban service provision as it correlates with urban infrastructure	■ Academics and NGOs can study the relationship between road infrastructure and how core urban services (water, sanitation, energy) are actually provided ■ Private sector actors can identify gaps in service provision as market opportunities
Land cover	Compare the ratio of impervious to pervious land cover within an urban area	■ Transportation planners can measure the amount of existing car-related infrastructure (e.g., parking lots) within a city to inform more efficient planning ■ Academics and NGOs can study potential correlation with impervious surface and urban heat island effect ■ Academics and NGOs can study potential flooding impacts related to impervious land surfaces
Disasters	Identify what types of land use have been affected by disasters, and how	■ Academics and NGOs can compare how land use and development decisions have affected impacts from natural disasters ■ Governments can identify which types of land use were most affected by a disaster and estimate number of affected people

Vacancy	Identify vacant or underutilized land within an urban area	■ Urban planners can identify currently vacant or underutilized land within a city to prioritize future development and improve overall urban land use efficiency
Development	Monitor how urban areas have been developed, and are being developed, in near real time	■ Academics and NGOs can study impacts and efficiency of different patterns of urban development ■ Private sector can forecast future development ■ Governments can identify nascent expansion to ensure proper provisioning of core urban services ■ Financiers can use historical growth details to guide investments ■ Funders can identify cities in the developing world that are rapidly expanding to prioritize investments
Density	Measure urban density within urban areas and among neighborhoods	■ Academics and NGOs can compare built-up space versus population density ■ Transportation planners and other service providers can use data for mass transit scenario planning ■ Employers can consider labor availability and access when evaluating potential business locations
Buildings	Measure building footprints and heights	■ Governments can identify data gaps within their existing cadastral maps ■ Governments can identify new potential areas for tax collection ■ Academics and NGOs can study the relationships between energy use intensity and building floor area ■ Academics and NGOs can measure building floor area by land use type ■ Energy planners can identify rooftop solar potential
Employment	Identify where urban residents live and work	■ Transportation planners can understand the spatial distribution of residences and jobs and use data for mass transit scenario planning ■ Academics and NGOs can compare patterns of work/life differentiation in urban areas ■ Governments can identify and prioritize areas for improved urban service provision
Infrastructure	Evaluate income levels using infrastructure (e.g., roofing) materials as a proxy	■ Academics and NGOs can study which building materials are indicative of higher versus lower income in various geographies and evaluate income levels ■ Governments can estimate demographic and socioeconomic features of housing and commercial districts at high resolution
Energy use and climate change	Assess and compare the energy and greenhouse gas impacts of urban development patterns and practices from the neighborhood to national scale	■ Cities and other governments working to improve resource efficiency, improve urban development, and address climate change

ENDNOTES

1. The Atlas of Urban Expansion. For more discussion, see section 2.3, "Data Sources."

2. The Atlas of Urban Expansion mapped the same set of cities circa 1990, circa 2000, and circa 2015.

3. Additional information at Digital Globe: https://explore.digitalglobe.com/Webinar-Building-Footprints-Ecopia-on-demand.html.

4. Additional information at Orbital Insight: https://orbitalinsight.com/use-cases/land-use-land-cover-change.

5. Landsat is the current name for a long-running American satellite program, whose stewardship has over time shifted between several federal agencies. As described by USGS, "Since 1972, Landsat satellites have created the longest continuously acquired space-based, moderate-resolution data archive. This joint USGS/NASA initiative supports worldwide remote sensing studies and helps land managers and policy-makers make informed decisions about our natural resources and the environment" (USGS n.d.). For more information, see https://www.usgs.gov/land-resources/nli/landsat or https://landsat.gsfc.nasa.gov/.

6. According to the European Space Agency, "Copernicus is the most ambitious Earth observation program to date. It will provide accurate, timely, and easily accessible information to improve the management of the environment, understand and mitigate the effects of climate change and ensure civil security...This initiative is headed by the European Commission in partnership with the European Space Agency" (ESA n.d.a). For more information, see https://www.esa.int/Our_Activities/Observing_the_Earth/Copernicus/Overview3.

7. For informal but exemplary discussions of expanding offerings and shrinking costs in cloud-based computing services, see https://www.stayclassyinternet.com/articles/investigating-AWS-pricing-over-time/ and https://medium.com/@retomeier/an-annotated-history-of-googles-cloud-platform-90b90f948920.

8. The model can be applied anywhere, but will likely struggle to characterize LULC in places that do not resemble any of the locations used as training data for the model. A model trained using data from lush coastal areas, for example, may have high success in similar littoral zones but will struggle to characterize a city in the desert. However, this problem can be alleviated either by expanding training datasets—creating a model using both coastal and desert data—or by creating a handful of models to specialize in different regions, climatic zones, etc.

9. Additional information at Atlas of Urban Expansion: http://www.atlasofurbanexpansion.org/.

10. Additional information at Descartes Labs: https://www.descarteslabs.com/.

11. The platform is openly accessible to anyone with an account, which is free to create. New users must request access to the data catalog via email. Various data products, such as imagery from commercial satellite constellations, may be available only to certain users with proper licensing.

12. Data for all cities can be downloaded at http://atlasofurbanexpansion.org/data.

13. Approximated by authors based on the number of 10-hectare samples and the size of the urban extent in a given city. The study area frequently includes a large hinterland far beyond the urban core, so this figure may understate the proportion of the "city" that was mapped.

14. For the purposes of the Atlas of Urban Expansion, each city's "urban extent" or "study area" is defined according to physical criteria related to density of development, without regard for administrative divisions or borders—though these may coincide, as often happens with rivers and other bodies of water. All locales fall within this conurbation. The use of physical characteristics to define the urban area in no way precludes the use of any resulting model to generate maps or analyses that are strictly reconciled to administrative (or any other) boundaries.

15. Detailed instructions and Python scripts for processing raw Atlas of Urban Expansion files into machine learning-ready archives are available within the project Github repository, at https://github.com/wri/UrbanLandUse/tree/phase-iii_release/aue-preprocessing.

16. For the purposes of this paper, "Sentinel-2 imagery" refers to the Level-1C product, which represents top-of-atmosphere reflectance. The Level-2A product, representing bottom-of-atmosphere reflectance, was not available globally at the time these experiments were conducted (ESA n.d.b). As such we did not investigate the relative merits of these different products.

17. It is of course possible to "upsample" imagery, that is, create a finer grid of pixels from an original, coarser raster. There are well-established techniques for performing this, which employ sophisticated mathematics to extract as much information as possible from the original image. However, there are clear limits to the effectiveness of these techniques, as the actual physical sensors only capture a finite amount of information. Whenever employed by this project, upscaling operations were performed by the Descartes platform. Details of this process are available at https://docs.descarteslabs.com/guides/raster.html#resampler.

18. Descriptions of platform functionality such as filtering and mosaicking can be found in the detailed documentation. For example, see https://docs.descarteslabs.com/guides/raster.html#mosaicking.

19. This imagery may be at its satellite's native resolution or at some higher (upsampled) or lower (downsampled) resolution. The process is the same regardless, with ground-truth vectors converted to rasters at the same resolution as the imagery sampling.

20. Each training sample is composed of an input and an output. For a given sample, the input is an array of numeric values, corresponding to the various bands of the pixels at and around the location of interest. The output is a single value representing LULC at that single location/pixel. This differs from the training data used in a semantic segmentation-type model, for example, where the output is effectively a grid of classifications congruent to the input look window.

21. As described by Harris Geospatial Solutions, in the context of remote sensing "Spectral indices are combinations of spectral reflectance from two or more wavelengths that indicate the relative abundance of features of interest" (L3Harris n.d.). For a brief introduction to spectral

indices, and an illustrative walkthrough of how to calculate them using geospatial software, visit https://ethiopia-gis.nrel.colostate.edu/pdf/RSLessons/RS_Lesson_6_Spectral_Indices.pdf.

22. Upsampling performed via bilinear interpolation and executed by the Descartes platform. More information is available at https://docs.descarteslabs.com/guides/raster.html#resampler. All bands were upsampled to a common resolution of 5m/pixel, despite having mixed native resolutions.

23. To be precise, these are bands 2, 3, 4, 8, 11, and 12, as illustrated in the official documentation here: https://earth.esa.int/web/sentinel/user-guides/sentinel-2-msi/resolutions/spatial.

24. The normalized difference vegetation index is probably the most well-known spectral index used in remote sensing. It is designed to distinguish chlorophyll-rich vegetation from other features. For an introduction to this index, visit https://gisgeography.com/ndvi-normalized-difference-vegetation-index/.

25. The use of spectral indices as artificial bands, or "pseudobands," for machine learning-driven LULC is a common technique, as demonstrated by Abdi (2019), Sun et al. (2019), and others.

26. We utilized gdal_rasterize, a function in the well-known Geospatial Data Abstraction Library (GDAL). "GDAL is a translator library for raster and vector geospatial data formats that is released under an X/MIT style Open Source License by the Open Source Geospatial Foundation" (OSGF n.d.). For more information see https://www.gdal.org/gdal_rasterize.html.

27. This means that the machine learning algorithm, seeking to establish patterns relating local appearance to LULC, could "see" at least 17 pixels—85 meters—in any direction from the central location; that is, the pixel the algorithm is trying to classify.

28. The details of the networks' composition—layers, structures, activation functions, and so on—can be viewed within the notebook at https://github.com/wri/UrbanLandUse/blob/phase-iii_release/notebooks/core_train-model-3category.ipynb.

29. Because the imagery utilized was calibrated to top-of-atmosphere reflectance, we trained all our models against a handful of different satellite captures, so that models would be exposed to various levels of image brightness, which varies with time of day, time of year, and atmospheric conditions. This, hopefully, would subsequently enable the trained models to more accurately classify input imagery with varying levels of brightness.

30. At this stage of research, we sought to simplify the modeling task to the extent possible, in order to establish a performance baseline—in essence, proof of concept. Creating a separate model for each city, rather than a unified and more complex multi-city model, was one part of this. Similarly, we asked each of those models to "learn" how a city looks during a single season. For example, a single tree-filled park may look very different with bare branches in winter, buds in the spring, dense green leaf cover in summer, and kaleidoscopic leaf cover of varying density in autumn. It is manifestly a more complex task to associate four or more signatures with a single LULC type than just one. Much of India is obscured by clouds during seasonal rains falling somewhere between May and October. Because clear imagery is critical for training, we avoided that time period entirely, and selected imagery from November, December, and January, in the hopes that lush vegetation would help the model distinguish between different types of LULC. With that said, limited testing suggested that the model would perform similarly if trained on imagery from before the rainy season, when the landscape is visibly dry across much of the country. These tests also indicated that models trained on imagery from one season performed better than models trained on imagery from throughout the year.

31. We sought to construct training data from imagery spanning as short a period of time as possible, in order to minimize discrepancies due to genuine, on-the-ground changes in LULC. For similar reasons, we also sought to use imagery captured as close as possible to the date of ground-truth collection, which occurred during or before 2015. The second satellite of the Sentinel-2 constellation became operational in mid-2017, effectively doubling the rate of capture, and thus the "density" of imagery over time. So we used imagery from late 2017 (and occasionally early 2018) to construct our training data for all Indian cities.

32. This preprocessing comprised normalization followed by removal of the mean and scaling to unit variance, as described here: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html.

33. Training data were divided on a random, per-sample basis, rather than via a geographic partitioning of the city—that is, done on a pixel-by-pixel basis, not locale-by-locale. This was a consequence of our workflow structure, but in the future we may transition to the latter, in order for the validation part of training to better approximate the application of the model to truly out-of-sample data.

34. Over the course of the project, we tested a number of modeling structures. In the pilot phase, we employed a stochastic gradient descent classifier (https://scikit-learn.org/stable/modules/sgd.html). Simple but nimble, this structure was appropriate for exploring and varying a wide variety of parameters, from imagery sources and spectral band combinations to more abstruse machine learning parameters. These models were implemented and trained using the popular open-source machine learning package scikit-learn (https://scikit-learn.org/). In subsequent phases, we tested other model structures of increasing sophistication, including fully connected neural networks. Eventually we progressed to convolutional neural networks (CNNs), a modeling structure that is highly popular for image-related tasks: "A [CNN] is able to successfully capture the Spatial and Temporal dependencies in an image through the application of relevant filters" (Saha 2018). At the cost of greater complexity and longer training times, CNNs have yielded our most accurate results to date. As our research continues, we will adjust and experiment with our model structures, but expect to remain within the CNN paradigm.

35. Keras library is housed at https://keras.io/.

36. TensorFlow library is housed at https://www.tensorflow.org/.

37. Because buildings are evaluated based on their appearance in satellite imagery, mixed-use buildings are typically classified according to their top floor and roof, which frequently results in a residential designation.

38. The structure of the "3-category model" was essentially identical to the structure of the "roads model," differing only at the end of the network, due to the different nature of the categorical and binary classification tasks. The loss functions for training each also differed for similar reasons. The most significant discrepancy between the two trainings was the use of two, subtly different sets of ground-truth data. Both were derived from identical AUE LULC polygons. However, for the training of the "roads model," that ground-truth was rasterized slightly differently: every grid element overlapping a road polygon, however incidentally, was classified as "road," rather than assigned the LULC type most prevalent in the pixel. The result was slightly wider road inputs for the road model, which was intended to improve performance in detecting smaller roads. This rasterization process can be viewed here: https://github.com/wri/UrbanLandUse/blob/4dc56f6c821e671d0b10d14a87e09916381cf541/utils/util_descartes.py#L154.

39. Code available at https://github.com/wri/UrbanLandUse/tree/phase-iii_release.

40. We calculated our F-scores using a beta value of 2 (an "F2 score"). A detailed description of the F-score and its calculation is available here: https://scikit-learn.org/stable/modules/model_evaluation.html#precision-recall-f-measure-metrics. The scores presented in the main results table represent the macro-average F-score, based on model performance on several discrete images of the study area.

41. For discussion of earlier results produced by models created using earlier configurations and different machine learning structures, please see Automating land use/land cover mapping of urban environments using satellite imagery, (Brumby et al. 2017).

42. In some cases, each "image" was in fact a mosaic of contemporaneous satellite images due to the size of the area concerned.

REFERENCES

Abdi, Abdulhakim Mohamed. 2020. "Land Cover and Land Use Classification Performance of Machine Learning Algorithms in a Boreal Landscape Using Sentinel-2 Data." *GIScience & Remote Sensing* 57 (1): 1–20. doi:10.1080/15481603.2019.1650447.

Angel, Shlomo, Alejandro M. Biel, Jason Parent, Patrick Lamson-Hall, and Nicolás Galarza Sánchez. 2016a. "Volume 1: Areas and Densities." In *Atlas of Urban Expansion—2016 Edition*, 2 vols. New York; Nairobi; Cambridge, MA: New York University; UN-Habitat; Lincoln Institute of Land Policy. https://www.lincolninst.edu/sites/default/files/pubfiles/atlas-of-urban-expansion-2016-volume-1-full.pdf.

Angel, Shlomo, Patrick Lamson-Hall, Manuel Madrid, Alejandro M. Biel, and Jason Parent. 2016b. "Volume 2: Blocks and Roads." In A*tlas of Urban Expansion—2016 Edition*, 2 vols. New York; Nairobi; Cambridge, MA: New York University; UN-Habitat; Lincoln Institute of Land Policy. https://www.lincolninst.edu/sites/default/files/pubfiles/atlas-of-urban-expansion-2016-volume-2-full.pdf.

Asher, Claire. 2019. "How Much Land on Earth Is Inhabited?" Curious Meerkat. http://www.curiousmeerkat.co.uk/questions/much-land-earth-inhabited/.

Banzhaf, Ellen, and René Höfer. 2008. "Monitoring Urban Structure Types as Spatial Indicators with CIR Aerial Photographs for a More Effective Urban Environmental Management." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 1 (2): 129–138. doi:10.1109/JSTARS.2008.2003310.

Brumby, Steven P., Peter Kerins, Eric Mackres, Emily Nilson, and Kate Owens. 2017. "Automating Land Use/Land Cover Mapping of Urban Environments Using Satellite Imagery." Presented at the 2017 IEEE Advanced Imagery Pattern Recognition Workshop, Washington, DC, October.

EEA (European Environmental Agency). n.d.a. "CORINE Land Cover — Copernicus Land Monitoring Service." Land Section. Copernicus. https://land.copernicus.eu/pan-european/corine-land-cover.

EEA. n.d.b. "Urban Atlas — Copernicus Land Monitoring Service." https://land.copernicus.eu/local/urban-atlas.

EEA. 2017. "Copernicus Land Monitoring ServiceLocal — Component: Urban Atlas." https://land.copernicus.eu/user-corner/publications/ua-flyer/.

ESA (European Space Agency). n.d.a. "Overview." https://www.esa.int/Applications/Observing_the_Earth/Copernicus/Overview3.

ESA. n.d.b. "Level-2A Processing Overview — Sentinel-2 MSI — Technical Guide — Sentinel Online." https://earth.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-2a-processing.

ESA. 2015a. "ESA CCI Land Cover Website." ESA Climate Change Initiative. https://www.esa-landcover-cci.org/.

ESA. 2015b. "Sentinel 1 — Data Access and Products." In *Sentinel Online Handbook*. https://sentinel.esa.int/documents/247904/1653440/Sentinel-1_Data_Access_and_Products.

ESA. 2015c. "Sentinel 2 User Handbook." In *Sentinel Online Handbook*. https://sentinel.esa.int/documents/247904/685211/Sentinel-2_User_Handbook.

ESA. 2017. "Color Vision for Copernicus: The Story of Sentinel-2." *ESA Bulletin* 161, 1st quarter. http://esamultimedia.esa.int/docs/EarthObservation/Sentinel-2_ESA_Bulletin161.pdf.

Gamba, Paolo, Fabio Dell'Acqua, Gianni Lisini, and Giovanna Trianni. 2007. "Improved VHR Urban Area Mapping Exploiting Object Boundaries." *IEEE Transactions on Geoscience and Remote Sensing* 45 (8): 2676–82. doi:10.1109/TGRS.2007.899811.

Graesser, Jordan, Anil Cheriyadat, Ranga Raju Vatsavai, Varun Chandola, Jordan Long, and Eddie Bright. 2012. "Image Based Characterization of Formal and Informal Neighborhoods in an Urban Landscape." *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 5 (4): 1164–1176. doi:10.1109/JSTARS.2012.2190383.

L3Harris Geospatial Solutions. n.d. "Spectral Indices." https://www.harrisgeospatial.com/docs/SpectralIndices.html.

NASA (National Aeronautics and Space Administration). n.d. MODIS Web. NASA. https://modis.gsfc.nasa.gov/about/specifications.php.

OSGF (Open Source Geospatial Foundation). n.d. "GDAL — GDAL Documentation." https://gdal.org/.

Patlolla, Dilip R., Anil M. Cheriyadat, Harini Sridharan, Vincent C. Paquit, Jeanette E. Weaver, and Mark A. Tuttle. 2013. "Mapping and Characterizing Global-Scale Human Settlements Using HPC." Poster presented at the International Conference for High Performance Computing, Networking, Storage and Analysis, Denver, CO, November 18. http://sc13.supercomputing.org/sites/default/files/PostersArchive/post240.html.

Saha, Sumit. 2018. "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 Way." Medium. https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53.

Sun, Chuanliang, Yan Bian, Tao Zhou, and Jianjun Pan. 2019. "Using Multi-source and Multi-temporal Remote Sensing Data Improves Crop-Type Mapping in the Subtropical Agriculture Region." *Sensors* 19 (10): 2401. doi:10.3390/s19102401.

USDA (U.S. Department of Agriculture). n.d. "NAIP Imagery." https://www.fsa.usda.gov/programs-and-services/aerial-photography/imagery-programs/naip-imagery/.

USGS (U.S. Geological Survey). n.d. "Global Land Survey (GLS)." https://www.usgs.gov/land-resources/nli/landsat/global-land-survey-gls?qt-science_support_page_related_con=0#qt-science_support_page_related_con.

USGS. 2019. "Landsat 8 (LS8) Data Users Handbook." https://prd-wret.s3.us-west-2.amazonaws.com/assets/palladium/production/atoms/files/LSDS-1574_L8_Data_Users_Handbook-v5.0.pdf.

Warren, Michael S., Steven P. Brumby, Samuel W. Skillman, Tim Kelton, Brendt Wohlberg, Mark Mathis, Rick Chartrand, Ryan Keisler, and Mark Johnson. 2015. "Seeing the Earth in the Cloud: Processing One Petabyte of Satellite Imagery in One Day." Paper presented at the IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington, DC, October 1–12. doi:10.1109/AIPR.2015.7444536.

ACKNOWLEDGMENTS

We are pleased to acknowledge our institutional strategic partners, who provide core funding to WRI: Netherlands Ministry of Foreign Affairs, Royal Danish Ministry of Foreign Affairs, and Swedish International Development Cooperation Agency.

The authors would like to offer acknowledgments and thanks to the following contributors and partners:

- Sam Brooks Hyde and Fabien Laurier at National Geographic Society
- Patrick Lamson-Hall, Professor Shlomo Angel, Nicolás Galarza Sánchez, and Alejandro Biel, the Atlas of Urban Expansion team of the NYU Urban Expansion Program at the Marron Institute of Urban Management
- The entire Descartes Labs team, with special thanks to Jeremy Malczyk, Sam Skillman, and Jay Carlson
- Our WRI colleagues Kate Owens, Anjali Mahendra, Raj Bhagat Palanichamy, V. Surya Prakash, Rejeet Mathews, Jorge Macias, Céline Jacquin, and Elleni Ashebir
- Reviewers Samantha Kuzma, Leslie Dewan, Gregory Taff, John Brandt, and Thomas Esch

ABOUT THE AUTHORS

Peter Kerins is a Research Analyst at the World Resources Institute. Contact: peter.kerins@wri.org.

Emily Nilson is the Resource Watch Product Manager at the World Resources Institute.

Eric Mackres is the Data and Tools Manager for Urban Efficiency & Climate at WRI Ross Center for Sustainable Cities.

Taufiq Rashid is a Research Assistant with the Office of Science & Research at the World Resources Institute.

Brookie Guzder-Williams is the Director of Data Science at the World Resources Institute.

Steven Brumby is a Senior Fellow at the World Resources Institute.

ABOUT WRI

World Resources Institute is a global research organization that turns big ideas into action at the nexus of environment, economic opportunity, and human well-being.

Our Challenge
Natural resources are at the foundation of economic opportunity and human well-being. But today, we are depleting Earth's resources at rates that are not sustainable, endangering economies and people's lives. People depend on clean water, fertile land, healthy forests, and a stable climate. Livable cities and clean energy are essential for a sustainable planet. We must address these urgent, global challenges this decade.

Our Vision
We envision an equitable and prosperous planet driven by the wise management of natural resources. We aspire to create a world where the actions of government, business, and communities combine to eliminate poverty and sustain the natural environment for all people.

Our Approach
COUNT IT
We start with data. We conduct independent research and draw on the latest technology to develop new insights and recommendations. Our rigorous analysis identifies risks, unveils opportunities, and informs smart strategies. We focus our efforts on influential and emerging economies where the future of sustainability will be determined.

CHANGE IT
We use our research to influence government policies, business strategies, and civil society action. We test projects with communities, companies, and government agencies to build a strong evidence base. Then, we work with partners to deliver change on the ground that alleviates poverty and strengthens society. We hold ourselves accountable to ensure our outcomes will be bold and enduring.

SCALE IT
We don't think small. Once tested, we work with partners to adopt and expand our efforts regionally and globally. We engage with decision-makers to carry out our ideas and elevate our impact. We measure success through government and business actions that improve people's lives and sustain a healthy environment.

Maps are for illustrative purposes and do not imply the expression of any opinion on the part of WRI, concerning the legal status of any country or territory or concerning the delimitation of frontiers or boundaries.