WORLD RESOURCES INSTITUTE

TECHNICAL NOTE

MAPPING URBAN LAND USE IN INDIA AND MEXICO USING REMOTE SENSING AND MACHINE LEARNING

PETER KERINS, BROOKIE GUZDER-WILLIAMS, ERIC MACKRES, TAUFIQ RASHID, AND ERIC PIETRASZKIEWICZ

ABSTRACT

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 based on medium-resolution (10–30 m/pixel) satellite imagery. The system and maps deploy the LULC taxonomy of the *Atlas of Urban Expansion—2016 Edition*: open, nonresidential, roads, and four types of residential space. We used supervised machine-learning techniques to apply this taxonomy at scale. Distinguishing between recognizable, clearly defined types of land use within a built-up area, rather than merely delineating artificial land cover, enables a huge variety of potential applications for policy, planning, and research. We demonstrate the training and application of machine-learning-based algorithms to characterize LULC over a large spatial and temporal range in a way that avoids many of the onerous constraints and expenses of the traditional LULC mapping process: manual identification and classification of features.

This document supersedes the previous technical note "Spatial Characterization of Urban Land Use through Machine Learning" (Kerins et al. 2020), and the methodology described here supersedes our previously reported techniques. In the initial research phase reported in that earlier note, we created distinct models for each of 11 cities in India, and used each of those models to create and

CONTENTS

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

Suggested Citation: Kerins, P., B. Guzder-Williams, E. Mackres, T. Rashid, and E. Pietraszkiewicz. 2021. "Mapping Urban Land Use in India and Mexico Using Remote Sensing and Machine Learning." Technical Note. Washington, DC: World Resources Institute. Available online at wri.org/publication/mapping-urban-land-use-in-india-and-mexico.

assess comprehensive LULC maps of the corresponding city. In this version, we refine and extend those techniques. We hugely expanded the scope of application of the resulting models, and herein describe the creation of several new LULC data products: a six-category, areal LULC map characterizing all of urban India as of late 2016; another comprehensive LULC map of urban India as of late 2019; a comparable six-category, areal LULC map of nine cities in Mexico as of late 2018; and a roads-only binary LULC map of those same Mexican cities, also as of late 2018. The models that generated these products were all created using the same methodology, with different training data sets as the only key distinguishing factor.

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. As discussed in section 1.3 "Existing Techniques and Products," the range of possible applications depends on the nature and resolution of the LULC data, which in turn depend on satellite imaging capability, image availability, and computing power.

1.1 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 most reliable way of mapping LULC is both straightforward and obvious: physically go to a location and record what is there. However, this presents obvious difficulties when trying to capture larger areas. Many mapping efforts thus turn to manual classification of satellite (or aerial) imagery, but this is a time- and resource-intensive process. Classification is also complex: how does one distinguish between distinct but similar types of residential land use, or categorize a building with retail at street level but residences on the upper floors? Furthermore, the 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, as different human analysts cycle through and classification practices evolve, deliberately or not.

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 Earth's surface. 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 https://oceanservice.noaa.gov/facts/lclu.html.
b. Additional information at https://ghsl.jrc.ec.europa.eu/.
c. Additional information at https://www.nass.usda.gov/Research_and_Science/Cropland/SARS1a.php.

1.2 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, manual 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 Earth's land surface that is covered by settlements, and accounts for 95 percent of human population, on an ongoing basis (Asher 2019). Additionally, an automated system requiring only satellite imagery as input could regularly generate new LULC maps based on new images, offering up-to-date monitoring and unlocking longitudinal studies.

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. This is of course highly subjective, as well as complicated: the extent to which the information distinguishes different types of land use—basically, the precision of its taxonomy—dictates what decisions it can help inform.

Whereas many global typologies simply aggregate all artificial land cover into a single "built-up" category, the classification system employed here focuses on distinguishing LULC within urban spaces. Distinction among types of urban land use opens up a range of potential applications for urban planning and research that would be otherwise unavailable. This classification scheme is discussed in section 2.5.1, "Ground-truth."

1.3 Existing Techniques and Products

Some businesses have recently begun selling information related to urban LULC. DigitalGlobe—formerly a privately owned satellite imagery provider based in Colorado, now part of Maxar—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 "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, as well as publicly available (EEA n.d.a). 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). OpenStreetMap 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.

1.4 The Emerging Opportunity

Several obstacles have historically impeded attempts to automate LULC classification of urban areas using satellite imagery. Although some publicly accessible imagery was 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.). Distinguishing land use, as opposed to land cover, is not practical at this resolution, at least within urban areas where the scale of the features is often much smaller. Contemporary commercial imagery offers superior spatial resolution, but its high cost and low spatial coverage often introduce major challenges. Because most commercial systems typically acquire imagery on demand, many areas are ignored or seldom captured, posing challenges for time-series analysis; compounding this, fewer satellite visits means fewer chances to acquire cloud- and shadow-free imagery. Ground-truth (see Box 2), 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, has posed additional technical hurdles.

But this situation has changed. Thirty-meter imagery from Landsat[5] satellites is now in the public domain. Furthermore, the European Union's Copernicus Programme,[6] in particular the Sentinel-2 constellation, provides regular global coverage and features improved spatial and spectral resolution over previous satellite programs, including several 10-meter bands. New machine-learning techniques have advanced 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.

Our research group has already demonstrated the viability of automated urban LULC mapping based on medium-resolution satellite imagery by using machine-learning techniques to create bespoke models for 11 cities in India, and then using those models to generate comprehensive maps of the corresponding cities at several points in time (Kerins et al. 2020). In this technical note, we report a revised and expanded methodology, which feeds larger volumes of training data into more sophisticated architectures, yielding more generalized models capable of credibly mapping urban LULC at the national scale.

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 2.4.1 Atlas of Urban Expansion and 2.5 "Training Data Creation."

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.

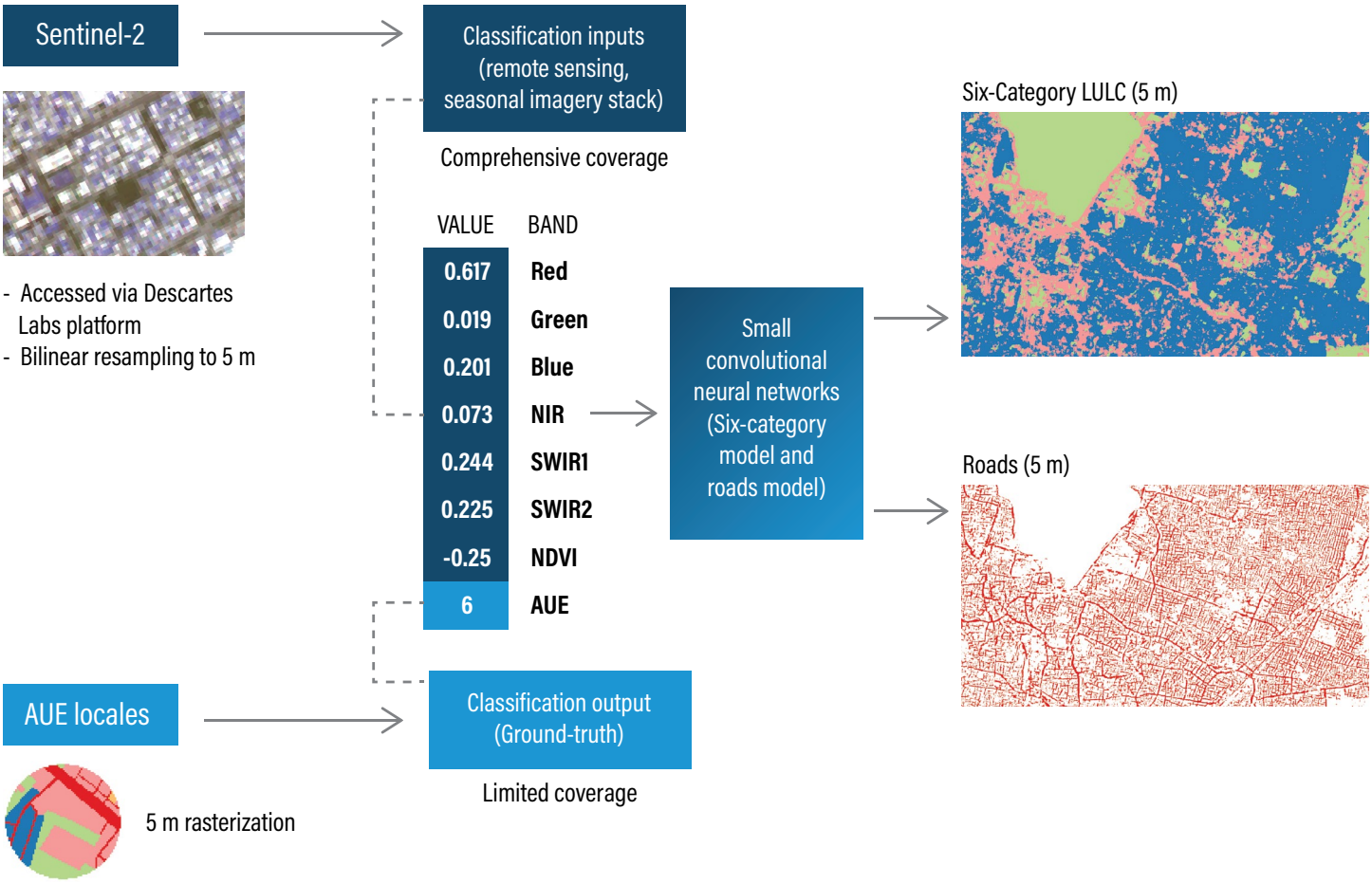
2. METHODOLOGY

The process for creating LULC classification models was straightforward but entailed many steps, from preparation of the computing environment to the processing of ground-truth to the construction of training data to the actual training of the models. Figure 1 attempts to summarize this workflow, which is described in detail in the following sections.

2.1 Requirements

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

Figure 1 | **Summary of Data Preparation and Modeling Workflow**



Note: Locations/pixels of known LULC are extracted from Atlas locales, and small squares of imagery, centered on those same locations, are extracted from Sentinel-2 images. These are combined into training samples, which pair the pixel values of the imagery with the classification value of the central pixel. Machine learning algorithms then use millions of these examples in order to associate certain "signatures" (i.e., features or patterns seen within satellite imagery) with specific LULC categories. Separate models are trained in this way for recognizing different areal types of LULC and for recognizing roads. Once trained, these models can classify LULC within new imagery, including places where the actual LULC is not known.
Source: Authors.

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 equations—that is, the model—in order to link inputs to expected outputs ever more closely. This process 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 2.5, "Training Data Creation"). Without both parts, the system cannot learn to associate various appearances with particular LULC categories—the crux of the project.

2.1.1 Scope and limitations

Once trained, a model can be applied to characterize LULC anywhere that imagery is available. Indeed, this ability to create comprehensive LULC maps extending beyond limited ground-truth coverage is the essential value proposition of this research. Because of the unrivaled temporal and spatial coverage of medium-resolution imagery, thanks to missions like Sentinel-2 and Landsat, our architecture is theoretically capable of generating truly global maps of urban areas across decades.

However, this is subject to a number of constraints. First of all, map quality hinges on the availability of relevant ground-truth. A trained model can be applied anywhere

but will struggle to accurately characterize LULC in images that do not resemble its training data. 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. Similarly, a model trained using images of a temperate city in summer will likely fail when applied to images of that same city in winter, because bare ground, leafless trees, or snow-covered roofs present novel, confounding appearances to the model.

In this research, we addressed these challenges by limiting the scope of any one model. Each model was trained using imagery and ground-truth from a single country (though a country like India presented highly diverse and occasionally confounding landscapes, as discussed in section 3.2.1 "India"). Furthermore, to reduce seasonal variability, models were trained with and applied to imagery only from late October to early January (in India, this represented the relatively cloudless post-monsoon period; in Mexico, the choice was not obviously significant).

By using different models—created from the same architecture and process, but with different training data—to characterize different regions or seasons, comprehensive LULC maps become feasible. (More generalized models trained using data from multiple regions or seasons offer an alternative pathway to coverage, though there is no bright dividing line between the two approaches—the India model, for example, is itself generalized for a variety of cities and landscapes in the subcontinent.) With that longer-term possibility in mind, we set out to test the feasibility of producing LULC classification and change-detection products that maintain fidelity and consistency across a diverse set of urban landscapes while using only publicly available data and open-source algorithms.

2.2 Computing Environment

Nearly all workflows, including data preparation, model training, and performance assessment, were implemented in Jupyter notebooks running a Python 3.6 kernel within a custom conda environment, and executed within those notebooks or standalone Python scripts—a common data science setup that offers flexibility and replicability.[8] The computing environment was Debian (Linux) within a virtual machine hosted on Google Compute Engine, built from a Google disk image for machine learning. Actual modeling was conducted using the Keras library[9] on top of TensorFlow.[10] Training and model application utilized a single Tesla K80 GPU.[11]

2.3 City Selection

The source of ground-truth for this project was the *Atlas of Urban Expansion—2016 Edition* (referred to as the Atlas). Early efforts suggested that our mapping algorithms struggled the most in the Indian subcontinent, relative to their performance elsewhere. For example, several simple support vector machine models yielded crude but reasonable-looking maps of some test cities like Budapest or Los Angeles but clearly failed in the denser urban fabric of Pune. Therefore, in order to develop the most robust and capable classifiers as we explored more sophisticated model structures, we focused on India for the research described in this paper. We employed the same processes and techniques to create models for Mexico, changing only the cities from which we drew training data. As with India, Mexico offered a relatively large amount of ground-truth as well as local knowledge of its cities from World Resources Institute (WRI) colleagues in the respective country office. Additionally, the stark difference between the two countries in urban layout and situation is immediately apparent in satellite imagery, allowing us to investigate how well our methodology might generalize to other regions. The 17 Indian and 9 Mexican cities are shown in Table 1.

Though these areas 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 "score"—that is, quantitatively assess—model performance in places where we have ground-truth data that allow us to compare model predictions to reality.

As discussed throughout this paper, the methods for India and Mexico differed slightly, though sample creation and model structure were identical. In India, more images were used for training data and the model was applied to images of all urban spaces across the country—several times in fact, as the methodology will show. The Mexico model was trained using only two images of each available city, and applied only across those same cities, and without that same level of geographic duplication. There were two main reasons for this discrepancy. The mapping in India was time-intensive and highly computationally demanding—in fact, not feasible for us if not for our partnership with Descartes Labs (see section 2.4.2, "Descartes Labs"). In Mexico, the mapping was conducted in response to project-related demand for preliminary

Table 1 | **Cities Represented in Training Data**

MODEL	INDIA		MEXICO
Cities represented within training data	Ahmedabad	Kozhikode	Culiacán
	Belgaum	Malegaon	Guadalajara
	Coimbatore	Mumbai	*León
	Hindupur	Parbhani	*Mérida
	Hyderabad	Pune	Mexico City
	Jaipur	Singrauli	*Monterrey
	Jalna	Sitapur	Reynosa
	Kanpur	Vijayawada	*Tijuana
	Kolkata		*Tuxtla

Note: These represent all Indian and Mexican cities for which LULC data were available from the Atlas of Urban Expansion team. An asterisk denotes a city that was not included in the original Atlas, in which case, we contracted the Atlas project and its team of imagery analysts to generate additional ground-truth data for these and several other cities.
Sources: Authors.

output, which was time-sensitive. Thus, we followed an abridged process to produce meaningful output for evaluation in Mexico. Nevertheless, the vast majority of the process was identical in both countries.

2.4 Data Sources

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

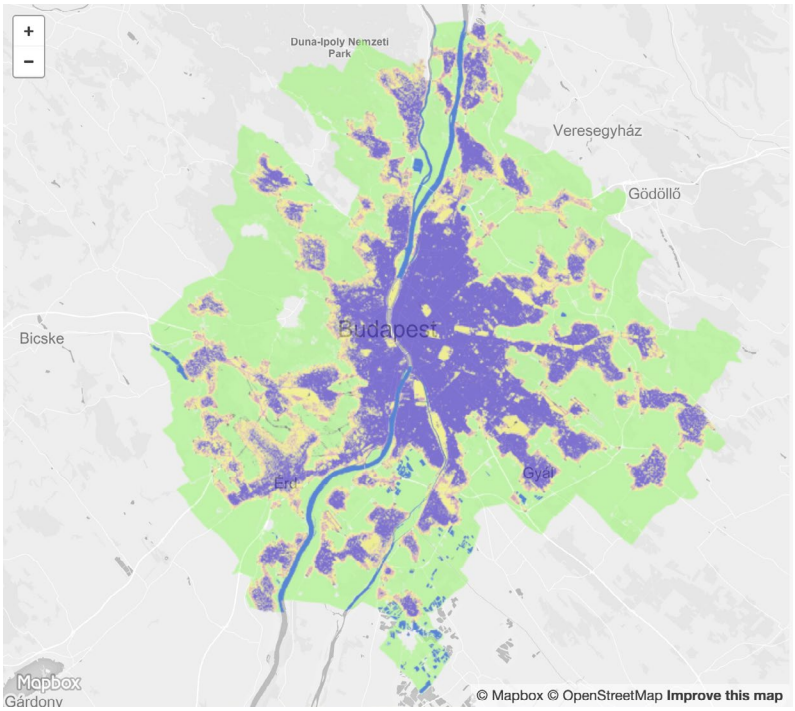
2.4.1 Atlas of Urban Expansion

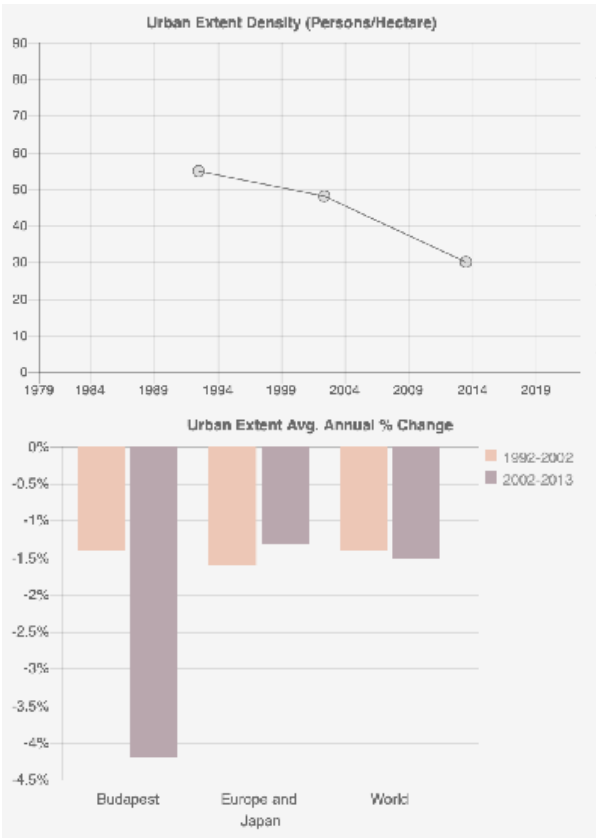
The Atlas of Urban Expansion[12] is a multiphase research effort conducted by the New York University (NYU) Urban Expansion Program in partnership with UN-Habitat and the Lincoln Institute of Land Policy. The Atlas, which analyzes a global sample of 200 cities, aims to examine how and why urban areas 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 supervised classification of medium-resolution Landsat imagery, followed by manual digitization of 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 (Figure 2).

To our knowledge, the Atlas represents the most detailed, best-documented, and most 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, because the data set is continental rather than global, and employs a much deeper set of categories, which 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.

Figure 2 | **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).

2.4.2 Descartes Labs

Descartes Labs[13] 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 data sets. Descartes Labs partnered with WRI to demonstrate its technology on a range of real-world applications with high potential impact for WRI's mission.

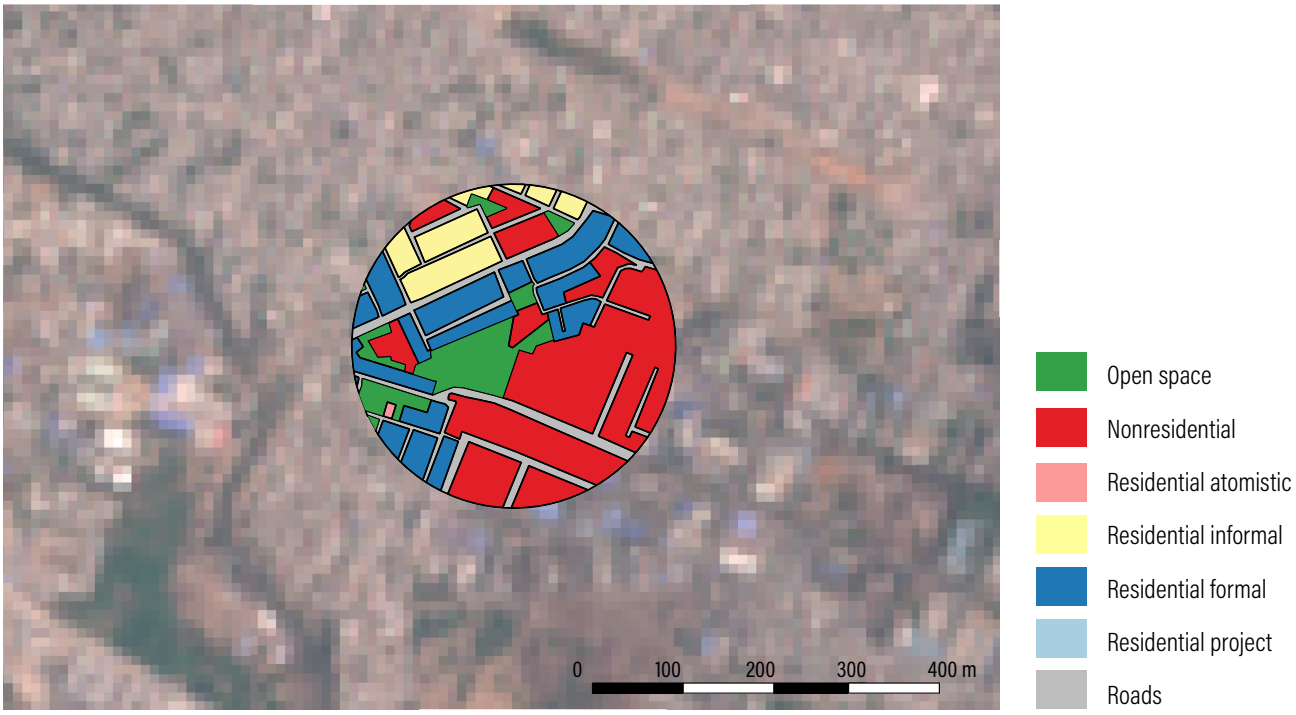
Descartes Labs has implemented a cloud-based software platform for daily ingestion, cleaning, and calibration of public-domain satellite imagery, including all data from Landsat, Sentinel-1, and Sentinel-2.[14] This platform for turning satellite observations into analysis-ready data is highly scalable, 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 satellite and aerial imagery, which our project executed via private compute nodes within Google Cloud Platform.

2.5 Training Data 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 that take as input the pixel values from satellite imagery at a given location, and return as output an LULC category (e.g., "residential"). We used supervised learning methods, in which a model "learns" from training samples, each of which links

Figure 3 | **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.

an example input to the corresponding, actual output. The process of model creation thus begins with training data construction.

2.5.1 Ground-truth

The Atlas of Urban Expansion project makes its data freely available[15] via a large number of geospatial files for each studied city. These contain data 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 data sets were essential to our work. One outlines the study samples, or "locales": 10-hectare circles quasirandomly selected from throughout each studied city, including both the urban core and the periphery, but not the large, often agricultural hinterland. Collectively, these circles cover roughly 1 percent[16] of an urban area.[17] This is substantial: the data set for a large city like Mumbai includes hundreds of these locales, constituting 20–30 square kilometers. The second data set consists of polygons contained within these locales. As shown in Figure 3, these polygons employ the Atlas's typology to classify LULC across the entirety of each locale, and we leveraged every square meter, whether as training or validation data. These polygons were manually created by a team of satellite analysis experts contracted by the Atlas 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.

As mentioned earlier, the typology expressed by the Atlas LULC data is crucial: discriminating between types of LULC within urban spaces is an essential, distinguishing feature of our work, and in large part defines its potential applications. Here we give an overview of the classification system; a thorough breakdown can be found in Atlas materials.[18] This is intended as a universal urban typology, so the categories rely on generally available aspects

of land use, rather than parochial traits or distinctions. "Formal" and "informal" refer strictly to physical features, not any legal status.

- Open Space includes open countryside, forests, cultivated lands, parks, vacant lands that have not been subdivided, cleared land, and water bodies: seas, rivers, lakes, and canals.
- Nonresidential Areas include all built-up areas, both public and private, that are not in residential use.
- Atomistic Settlements are areas with irregular layouts that were clearly not subdivided or laid out before residential construction took place. This category includes squatter settlements that grew incrementally without an overall plan, homes built on irregular parcels of land, or homes built on rural plots that were not regularly subdivided before their conversion to urban use.
- Informal Land Subdivisions are areas that have been subdivided for urban use but that lack visible evidence of conformity to land subdivision regulations such as regular plot dimensions, paved roads, streetlights, or sidewalks. That said, structures in these informal land subdivisions, although different in size and form, are typically laid out along straight or almost-straight roads, with regular intersections and standardized widths. Blocks are also regular or semiregular in size and shape, when topography permits.
- Formal Land Subdivisions are similar in layout to informal layouts, but exhibit a higher level of regularity, a higher level of provision of infrastructure, and better connections to existing roads. All roads must be paved for an area to be classified as a formal land subdivision. Sidewalks and streetlights are often visible as well.
- Housing Projects range from large apartment tower projects to suburban tract housing. Housing projects share one feature: their structures must be essentially homogenous. These are projects in which all structures are built by a single developer using variations on the same plan.
- Road Space includes the rights-of-way of lanes, streets and roads, both paved and unpaved, containing both the area that is currently in use and any lands that are clearly reserved for future use (Angel et al. 2016b).

These polygons are the foundation of our ground-truth: they explicitly ascribe LULC to precisely demarcated areas based on morphological criteria. However, preparing them for machine-learning ingestion required a significant amount of processing.[19] 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 an LULC category. The final step, in which these vectors were converted into raster data, is described in section 2.5.3, "Combining imagery and ground-truth."

2.5.2 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 2 summarizes these differences, while Figure 4 gives a detailed breakdown of the imaging capabilities of Sentinel-2, our chosen source of satellite imagery.[20] Figure 5 provides an example of the difference in spatial resolution between imagery sources, including the nonsatellite National Agriculture Imagery Program (NAIP), which conducts aerial photography. We did not investigate the use of nonimagery inputs, such as digital elevation models, OpenStreetMap vectors, or image capture metadata (e.g., view angle of satellite or time of day).

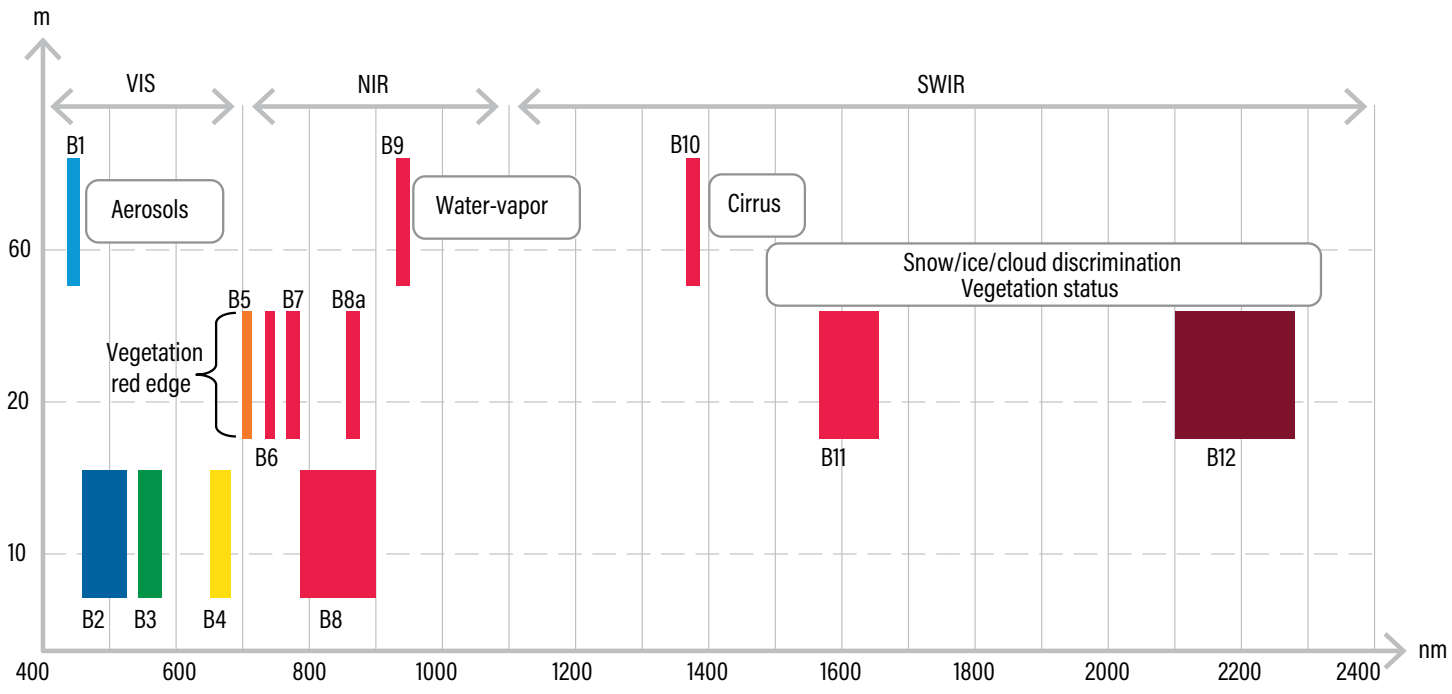
Spatial resolution was particularly significant for a project focused on fine urban features like roads. Our images were rasters: essentially, grids of pixels. The spatial resolution of the imagery determines the size of the grid elements, and thus the area of Earth's surface captured by each pixel. 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 an LULC to it, we refer to a pixel, and the entire area it covers. As described later, we upsampled our imagery to generate "new" pixels at a resolution higher than the original sensor data.[21] Those upsampled pixels

Table 2 | **Comparison of Selected Public Remote Sensing Systems**

SATELLITE	LANDSAT	SENTINEL-1	SENTINEL-2	NATIONAL AGRICULTURE IMAGERY PROGRAM (NAIP)
Spatial resolution	15–30 m/pixel	20 m/pixel	10–20 m/pixel	1 m/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 4). 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, 2015c). Metadata for NAIP from USDA (n.d.).

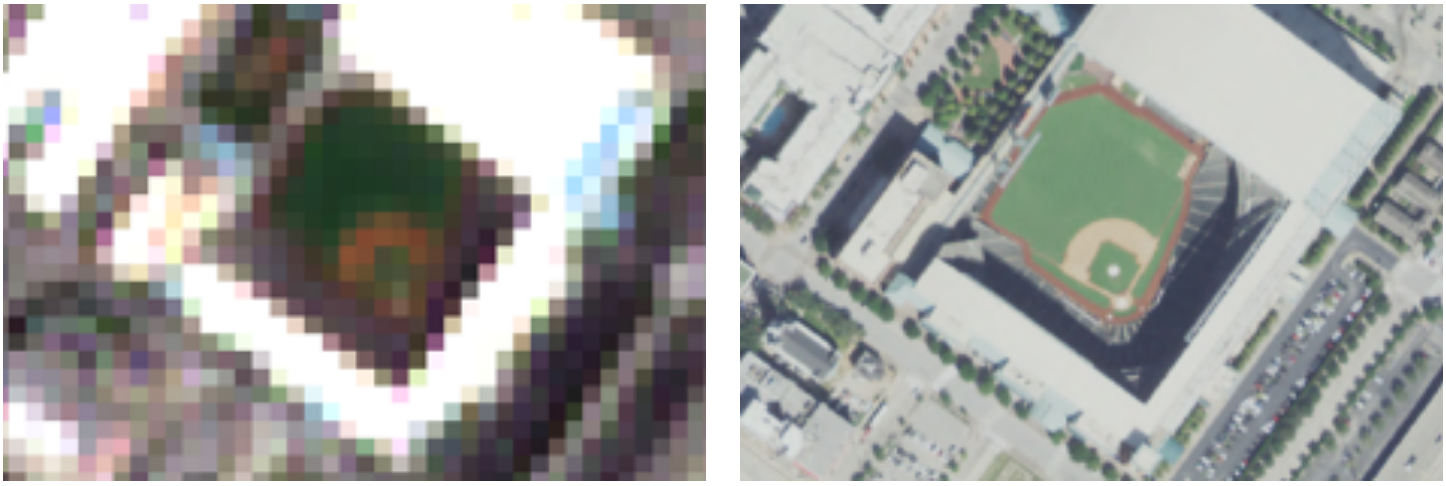
Figure 4 | **Imaging Capabilities of Sentinel-2**



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 5 | **Comparison of Sentinel-2 and National Agricultural Imagery Program (NAIP) Imagery**



Note: On the left, Sentinel-2 10 m/pixel imagery of the baseball stadium in downtown Houston; on the right, the same ballpark as captured by NAIP at 1 m/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.

represent the limit of the precision of our analysis, and we did not evaluate intrapixel variation, though of course any pixel may in reality contain multiple types of LULC.

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. For example, Sentinel-2's combination of high revisit rate 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, as follows. 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.[22] 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.[23]

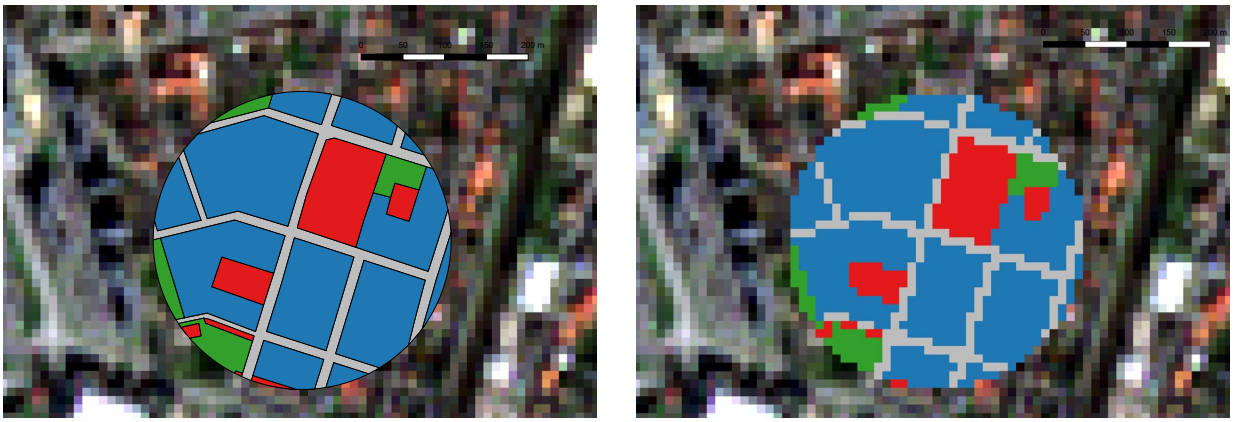
Ultimately, we used the Sentinel-2 Level-1C product, which represents top-of-atmosphere reflectance (ESA n.d.b). We did not "composite" or otherwise combine images, with the exception of mosaicking multiple images when necessary to get a picture of the full study area. In these cases, the mosaic constituents were always contemporaneous, typically captured on the same day. This merits emphasizing, as many comparable efforts use composite or "representative" images: training data were constructed directly from actual satellite captures, and not any sort of annual or seasonal composite or "median" image. This choice was motivated in part by a desire for the system to be able to immediately classify new images as they are captured.

2.5.3 Combining imagery and ground-truth

Our machine-learning algorithm characterizes the LULC at a location—that is, at a single pixel—by examining the imagery at and around that point, an area known as the "look window." This local imagery is expressed via a three-dimensional numeric grid, whose individual entries correspond to the brightness in each spectral band for each pixel within the look window. Such a grid preserves spatial relationships between different locations within the imagery, which allows a model to learn and leverage spatial patterns. For a rudimentary example, the mere presence of several gray pixels in an area might not seem meaningful, whereas their appearance in a neat row, flanked by nongray pixels, might well indicate a road.

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

Figure 6 | **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 constitutes one training sample. Using 5 m/pixel imagery, each 10-hectare locale constituted 40,000 pixels.
Sources: Underlying imagery from Sentinel-2 accessed via the Descartes Labs platform; LULC polygons from the Atlas of Urban Expansion; graphic by authors.

of the underlying imagery,[24] as illustrated in Figure 6. 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 three-dimensional numeric grid representing the look window around a pixel is joined with the known LULC classification of that location, that combination constitutes one "sample."[25] (When extracting the sample input from imagery, we used a 17 x 17 square of pixels centered around the LULC site, resulting in a look window 85 meters wide).[26] When repeated for every pixel with a known LULC classification, we have a set of inputs and the corresponding desired outputs: our training data.

2.5.4 Our training data configuration

All results reported in this document, for both India and Mexico, were produced using a single, consistent configuration. 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 using bilinear interpolation, in hopes of utilizing as much information as possible from our high-precision ground-truth.[27] Training sample inputs comprised the following bands: blue, green, red, near infrared (NIR), shortwave infrared 1 (SWIR1), and shortwave infrared 2 (SWIR2).[28] Ground-truth vector files were transformed into geospatial rasters precisely coterminous with the satellite imagery rasters using the Geospatial Data Abstraction Library.[29]

These two sets of coterminous rasters—imagery and ground-truth—were then used to construct training samples. For each pixel with a known LULC classification, a small image centered around that pixel—an image "chip"—was extracted from the downloaded imagery and saved to file. An entry was then created in a master catalog to record the relevant features of each chip, such as file path, imagery source, resolution, and the actual LULC class of the central pixel (Table 3). Note that this process was repeated for multiple images of every city, meaning that the catalog contained multiple image chips corresponding to each pixel of ground-truth. Collectively, these represent the appearance of that location at different times and under various conditions during the time period of interest. The catalog itself was accessed and edited as a pandas[30] DataFrame object, and stored as a simple comma-separated values file. This catalog-based approach allowed us to precisely filter entries with ease.

2.6 Training Data Set Compilation

A collection of samples used to train a model is a training data set. Compiling such a set was a matter of constructing an appropriate chip catalog. This in turn was a matter of filtering the master catalog to exclude all but the desired entries. This primarily entailed ensuring that

Table 3 | **Entries from Chip Catalog**

#	PATH	CITY	LOCALE	SOURCE	IMAGE	BANDS	RESOLUTION	RESAMPLING	TILE_ID	X	Y	LULC
7	/data/phase_iv/sitapur/chips/none/sitapur_aue0...	sitapur	43	s2	E	['blue', 'green', 'red', 'nir', 'swir1', 'swir...	5	bilinear	176	254	173	0
8	/data/phase_iv/sitapur/chips/none/sitapur_aue0...	sitapur	43	s2	E	['blue', 'green', 'red', 'nir', 'swir1', 'swir...	5	bilinear	176	255	173	0
9	/data/phase_iv/sitapur/chips/none/sitapur_aue0...	sitapur	43	s2	E	['blue', 'green', 'red', 'nir', 'swir1', 'swir...	5	bilinear	176	250	174	1
10	/data/phase_iv/sitapur/chips/none/sitapur_aue0...	sitapur	43	s2	E	['blue', 'green', 'red', 'nir', 'swir1', 'swir...	5	bilinear	176	251	174	1

Note: This figure shows a truncated excerpt from the chip catalog, with several columns omitted. Highlighted in orange is the critical piece of information: the actual LULC classification of the pixel at the center of the corresponding image chip. On average, a city contained approximately 400,000 total samples (per image).
Source: Authors.

only selected images from selected cities were included. Again note that we created multiple training samples corresponding to each pixel of ground-truth; the distinct images used to do so show each location under various atmospheric conditions, times of day, and so on. Each training data set for the final models included multiple samples corresponding to several distinct images of every location. In some cases, this was followed by additional filtering. For example, to train a model intended to identify areal types of LULC—that is, all categories in our typology except for roads—we would exclude catalog entries corresponding to road LULC.

2.6.1 Excluded samples

Review of early results revealed a trend, pronounced in some cities and negligible in others, of models mislabeling open space as nonresidential land. Careful inspection revealed that these models were accurately reflecting the training data: in some cases, apparent open space was labeled as nonresidential land in the ground-truth. For example, at an airport, not only would the runway and terminals be labeled as nonresidential but so would the entire area of the airport, including huge areas of scrubland or grassy fields. While it is perfectly reasonable to consider the entire airport a commercial space in most contexts, such as zoning, from an LULC perspective it is confounding to label wide swathes of natural land cover as anything other than open space. Accordingly, we manually reviewed the ground-truth for such instances, and recorded the locales containing such confoundingly labeled areas, so that we could exclude them from training and validation data when desired. It is possible that the ground-truth contains additional types of confounding labels, such as mixed-use residential space. However, we were unable to reliably assess the "quality"—that is, the appropriateness for our purposes—of the LULC ascriptions, beyond the most obvious case of "mislabeled" open space.

2.6.2 Image selection

For the India models, training imagery was drawn from the post-monsoon period in a single year: October 2017 through January 2018. For simplicity, we used approximately the same time window for the Mexico model, extending through March. Images were selected in a two-step process. Starting with all Sentinel-2 scenes captured during the target date range, we first used the Descartes Labs platform to automatically filter out very cloudy images, as measured by the native cloud mask within the Level-1C product. We then manually reviewed the remaining images, looking in particular for obscuring clouds or their shadows over areas containing ground-truth and for washed-out coloration indicative of atmospheric haze. This process continued until we had approximately half a dozen clear, comprehensive images (in some cases, mosaics) of each city. These were the images from which

chips were extracted, catalog entries made, and which, when joined with our ground-truth data, comprised our pool of samples.

2.6.3 Training and validation tranches

After constructing an appropriate catalog of samples, we divided it into two subcatalogs: one containing chips to be used for training and the other, chips for validation. Chips were divided on the basis of locale: for each city, approximately 70 percent of locales (and thus 70 percent of samples) were randomly assigned for training, and the remainder assigned for validation. This assignment was permanent: samples drawn from every image of a city were divided the same way, and this separation was maintained for every training. Thus no training sample was ever used for validation, and vice versa. (Later in this paper, we also discuss "validation" scores in reference to these reserved locations, which were not used for training, rather than specific samples, in order to directly compare various interim and final LULC maps.) Furthermore, because the locales represent contiguous areas and are dispersed across an urban area, the look window of a training sample generally did not allow a model to "peek" into the validation data. Basically, this process geographically divided each city into training and validation areas, and this segregation was maintained for all model development.

2.7 Model Creation

"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 samples in a training data set, constructed as described earlier. 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 a 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.

2.7.1 Model structure

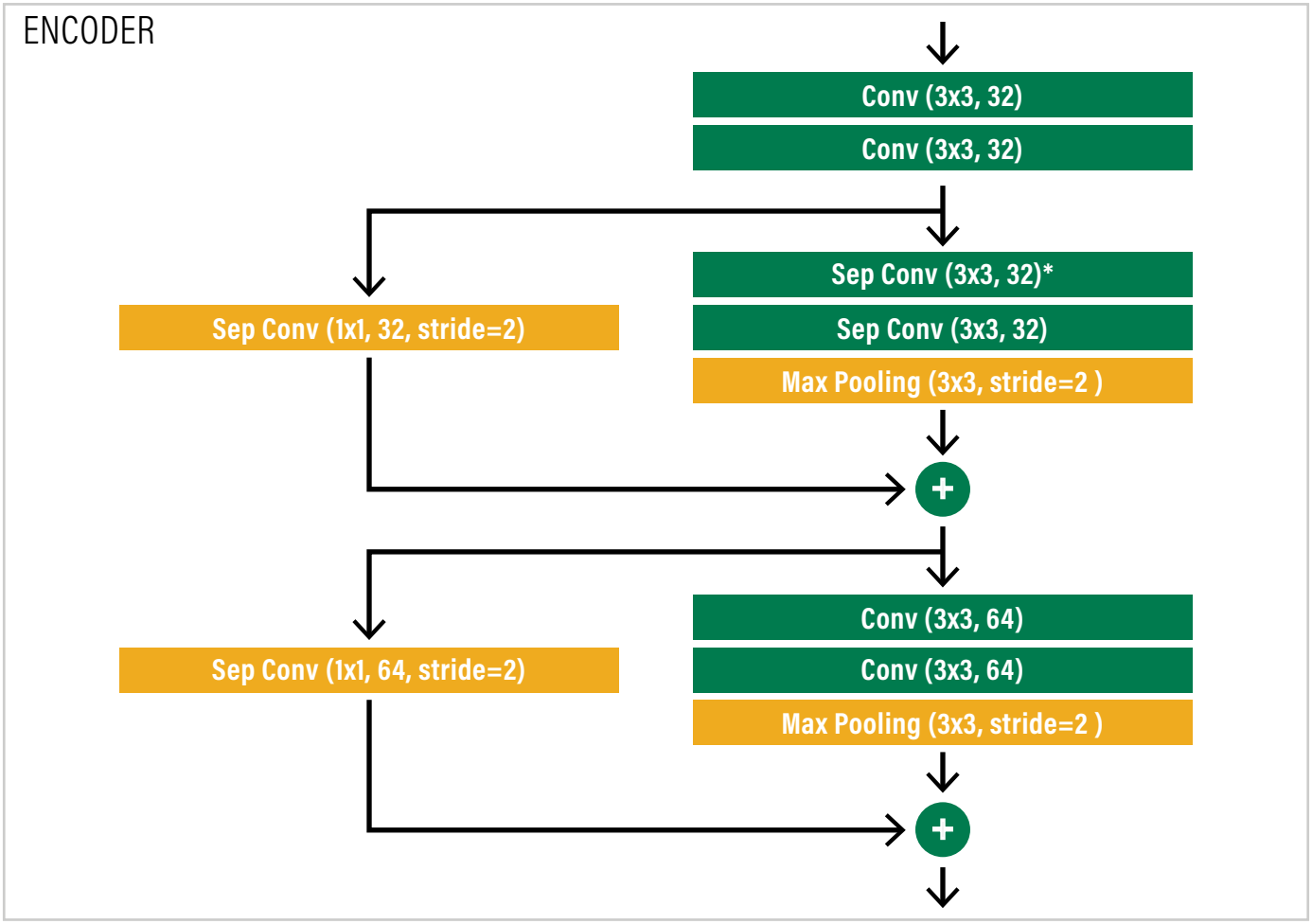
All models were small convolutional neural networks. The areal and roads networks were structurally identical, save for the number of output nodes in the final layer of each.[31] There was also no structural difference between the respective areal and roads networks for India and Mexico. The encoder for our models is shown in Figure 7. The architecture is based on the entry flow of the Xception Network (Chollet 2017). This structure has been reported as effective in remote sensing-based complex land cover classification (Mahdianpari et al. 2018). In particular, we make use of Separable Convolutions to give us a lighter-weight model that avoids overfitting. Each Conv Block contains a convolution followed by ReLU, repeated twice. The Separable Conv blocks contain a ReLU, followed by a Separable Convolution, followed by a BatchNorm Layer—also repeated twice. Note, however, that the first Separable Conv block (denoted with an asterisk) drops the initial ReLU. This encoder is then followed by a simple classifier containing two dense layers as shown in Figure 8.

2.7.2 Model training

Training was conducted by passing a chip catalog, constructed as described in section 2.5.4, "Our training data configuration," to a custom software module that used the list of chips to construct the actual training sample objects (paired input and output numeric arrays), compiled them into batches (sets of samples for model ingestion), and then passed those batches to the actual training routine.

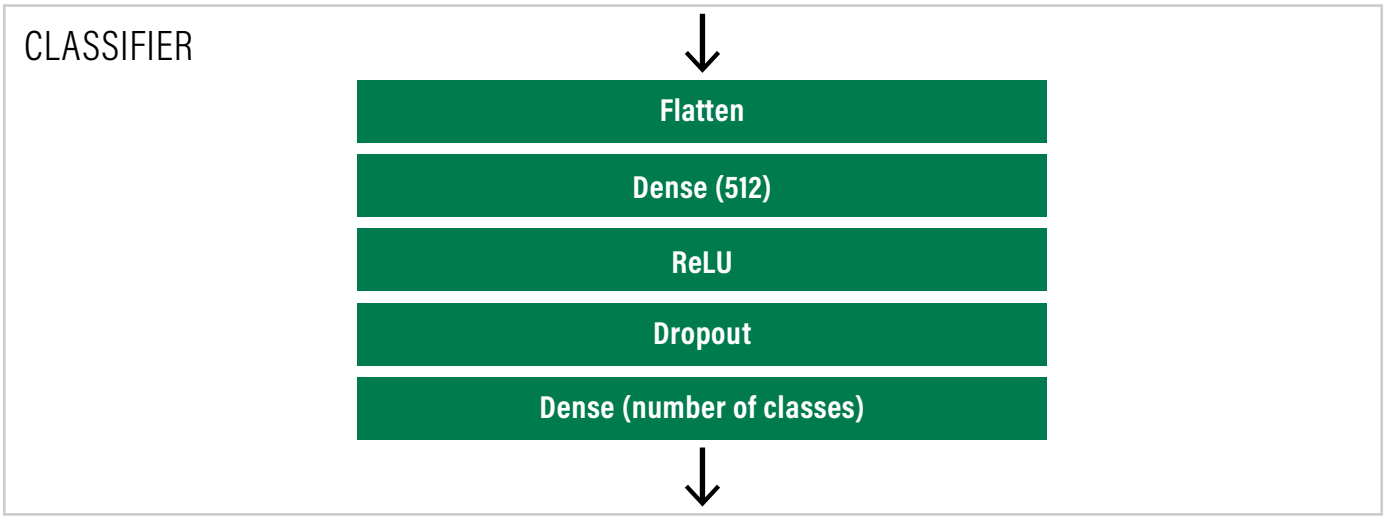
We were able to train models using very large data sets by using the *_generator* routines in Keras, which hold only a small fraction of the data in memory at any single moment. In previous iterations of research, the size of training and validation data sets was strictly limited by the amount of memory available. In theory, this new arrangement could allow training data sets of nearly unlimited size; in practice, Keras (as deployed within our computing stack) struggled when the size of data sets reached tens of millions of samples. This limitation did not preclude us from successfully training models, but with larger data sets we did have to implement some work-arounds, such as manually handling multiple training epochs, rather than allowing callback functions to manage the process.

Figure 7 | **Schematic of Model Encoder**



Source: Authors.

Figure 8 | **Schematic of Final Model Layers**



Source: Authors.

Our training infrastructure also allowed us to aggregate different LULC categories when desired. For example, although we do not present the results here, we trained several "three-category" models distinguishing only between open space, nonresidential, and "residential," where the last category was in fact composed of four different types of residential LULC as recorded in the Atlas. When we created models for mapping roads, we used a binary classification: "road" or "not road," where the latter includes every category in our typology besides road. Of course, we could also choose not to aggregate categories and train models fully leveraging the typology; this was the case for the areal models that generated the unaggregated, six-category maps.

All model trainings used a custom, weighted categorical cross-entropy loss function, which gives greater weight to the less-represented classes in the training data set.[32] Every model underwent two trainings: a "fast" training with a higher learning rate, and a "slow" training with a lower learning rate. Generally, the number of epochs varied inversely with the number of samples.

MODEL VALIDATION

As described in section 2.6.3, "Training and validation tranches," we reserved approximately 30 percent of samples for validation. In some machine-learning configurations, the model would be applied to these validation samples after every epoch, and training would continue until model performance against these validation data failed to improve. However, given the huge volumes of training data, we configured our machine-learning routines to last a fixed number of epochs. This means these were not used in the training process—allowing them to serve as true, out-of-sample "validation" data.

2.7.3 Our machine-learning configuration

We have outlined our general methodology, and the capabilities of our software infrastructure to train different models under various conditions. Here, we describe the training data and configuration used to create the models used to generate the data products presented in this paper. The areal and roads models for India and for Mexico were created and trained using the exact same process. The only difference lay in the training data sets.

TRAINING DATA

The India models were trained using samples from all 17 Indian cities in the Atlas. The number of images per city ranged from two to seven, based on the availability of clear imagery within the target date range, with a median of five per city, and a total of 85 images. Ultimately, the training and validation data comprised approximately 20 million samples. The Mexico models were trained using samples from nine Mexican cities. We used two images from every city, for a total of 18 images. Together, the training and validation samples numbered about 4 million. In all cases, we excluded all locales containing confounding LULC assignations, as described in section 2.6., "Excluded samples."

TRAINING CONFIGURATION

The models were trained as described above, using a custom loss function and a batch size of 128. Given the huge volume of samples, training entailed just two epochs: one "fast," with a learning rate of 0.001, and one "slow," with a learning rate of 0.0001. During each epoch, the model was exposed to every training sample once, in a random order. Weights were saved after each epoch, and the entire model object was saved at the conclusion of training.

SIX-CATEGORY CLASSIFIER

The areal LULC classification models did not aggregate any categories, so each sample simply represented the native classification of that pixel. However, we excluded from their training data sets all samples labeled as road. The areal models were trained as six-category classifiers, which could assign to a given pixel any of the LULC categories from the Atlas besides road.

ROADS CLASSIFIER

The sets of samples used to train the roads models did not exclude any samples on the basis of LULC classification; instead, all samples from categories other than road were aggregated to a single "not road" category. Thus, the resulting binary classifiers could simply identify a pixel as road, or not.

2.8 Model Application

To generate maps of cities and other limited areas, we simply used our virtual machine to apply trained models to input imagery. This was the process for Mexico, where we mapped the urban extent of nine cities. For each Mexican city, we simply applied the model to a single set of contemporaneous satellite captures which collectively covered the full urban extent (as defined by the Atlas). The resulting, partially overlapping LULC model outputs were then combined into a final LULC product as described in section 2.8.4, "Post-processing." In the case of India, though, the process was much more extensive.

2.8.1 Mapping at scale

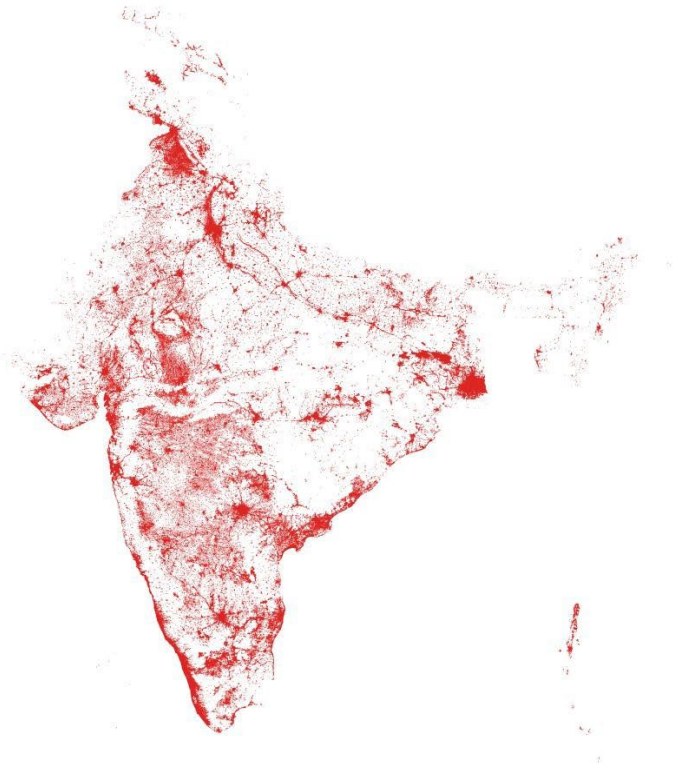
To apply the model at scale in India, we utilized the cloud computing infrastructure and tasking functionality of the Descartes Labs platform.[33] In essence, this combination of hardware and software allowed us to create hundreds of ephemeral virtual machines, and pass to each an identical copy of the model and a small area to map. Once each of these computing instances completed its mapping, it saved the resulting output to Descartes Labs storage and then terminated.

2.8.2 Area of application

Because the training data derived from the Atlas covers only cities, the resulting model is intended as a classifier only for urban areas. We therefore attempted to apply the model only to those areas across India. Rather than define that ourselves, we used the Global Human Settlement Layer (GHSL) to generate a mask of densely populated areas within India, and used that as a proxy for urban spaces. Specifically, the urban mask we created for India includes any land lying within the official Government of India national boundary that is identified in the GHS-BUILT 2014 1 km layer (Corbane et al. 2018) as at least 1 percent built-up. An additional buffer of 600 meters was added around every built-up area, in order to map as much of the total urban space as possible (at the cost of including more nonurban space on the fringes). We then used that mask (Figure 9) to generate a very large set of square tiles, each 512 pixels (2,560 meters) wide, covering all of urban India. If any part of a tile overlapped with the urban mask, LULC was predicted for the entirety of that tile.

It should be acknowledged that populated areas are not identical to urban areas, and the GHSL unavoidably includes some nonurban population centers, which may differ meaningfully from the large cities of the Atlas. Because we lacked ground-truth for nonurban

Figure 9 | **Map of "Urban India"**



Source: Underlying data from GHSL; processing and graphic by authors.

spaces, we could not quantitatively assess model performance in these areas, and output should be interpreted with caution.

2.8.3 Image selection

For each tile to be mapped, we used the Descartes Labs platform to identify all intersecting Sentinel-2 scenes captured within the target date range (November 1 through January 15 in the chosen year[34]). All such scenes for the 2016 and 2019 mappings represented novel imagery, as all training images were drawn from the 2017 season. Out of that set, the five scenes with the lowest level of cloud coverage[35] within that tile were automatically selected for mapping. Thus, for each tile constituting urban India, the model generated five independent LULC maps based on the five different satellite captures of the area.

2.8.4 Post-processing

After applying the model to all of urban India, we had five distinct LULC maps for every tile in our application area. Each of these direct model outputs was to some extent contingent on the conditions present at the time the corresponding Sentinel-2 scene was captured. Clouds, cloud shadows, atmospheric haze, smoke, and time of day are just some of the factors that introduce variability into a captured image. To mitigate the impact of these transient effects, we mapped each area on the basis of several independent input images, as described, and then combined them together into a final, composite output.

Compositing was conducted in a straightforward manner. For every pixel, each output map represented one "vote." The LULC category with a plurality of votes was ascribed to that pixel in the final composite map. For pixels that were obscured by clouds in a given scene, the corresponding pixels in the model output were omitted from the voting process. Given the nature of the compositing process, we refer to this final resulting map as a "mode map." The impact of this compositing on performance is discussed in section 3.1, "Scores," specifically, section 3.1.5, "India scores comparison."

As mentioned, this same process was employed with the Mexico LULC outputs. However, there the impact was limited. For each Mexican city we applied the model to a single set of satellite images that collectively covered the city, whereas in India we had five "comprehensive" mosaics for each tile to be mapped. As a result, in Mexico, multiple "votes" occurred only where satellite captures overlapped, and even in those instances, there were often only two relevant images.[36] In contrast, every mapped pixel in India had five independent votes (excepting those excluded due to cloud cover).

3. RESULTS AND DISCUSSION

In the context of India, the final result of our process was an areal LULC map covering the urban extent of the country. In Mexico, it was a more limited mapping of several cities, with separate areal and roads products. It was important to pair these outputs with quantitative assessments of accuracy. Because the maps extend far beyond the ground-truth—which was of course the entire purpose of the project—we estimated overall quality based on the limited areas where ground-truth was available. In each case, this consisted of comparing LULC predictions with the rasterized ground-truth, pixel by pixel. Furthermore, to ensure a legitimate estimate of quality, we limited the comparison to reserved validation data: pixels of known LULC that were not used in training the model, or at any subsequent stage of output creation. All presented statistical metrics reflect exclusively the same set of withheld validation locations (see section 2.6.3, "Training and validation tranches").

As described in section 2.8, "Model Application," the creation of our final LULC maps involved several steps. To better evaluate this multistage process, we quantitatively assessed—"scored"—outputs from several stages of the mapping process: direct application of model to the training images, direct application of model to novel imagery, and combination of multiple model outputs from multiple novel images into a final composite map.

3.1 Scores

For each product and at each step, we compared LULC classifications to reserved validation data in order to generate a confusion matrix.[37] We used the confusion matrix to calculate recall and precision—also known as producer's and user's accuracy, respectively—by category, from which F2 scores[38] were then derived. The F2 score summarizes performance in a single number while attributing twice as much significance to recall as to precision,[39] and we used it as our primary performance metric.[40] Like accuracy, the F2 score ranges from 0 to 1, with 1 signifying perfect performance.

Classifying LULC for samples that the models had never "seen" before revealed to what degree the models simply reflected the training data, and to what degree they

Table 4 | **Example Confusion Matrix of Composite Map for Hyderabad, 2019**

LULC CATEGORY MODEL		MODEL					
		OPEN SPACE	NONRESIDENTIAL	RESIDENTIAL - ATOMISTIC	RESIDENTIAL - INFORMAL	RESIDENTIAL - FORMAL	RESIDENTIAL - PROJECT
Ground-truth	Open Space	62,866	9,263	131	9,640	4,927	1,276
	Nonresidential	15,149	28,810	1,535	8,668	14,818	1,128
	Residential - Atomistic	355	1,149	182	4,877	3,231	22
	Residential - Informal	1,000	1,339	131	16,280	12,147	122
	Residential - Formal	786	1,864	357	17,486	53,374	25
	Residential - Project	342	74	15	94	393	1,092

LULC CATEGORY MODEL		MODEL		
		OPEN SPACE	NONRESIDENTIAL	RESIDENTIAL - ALL
Ground-truth	Open Space	62,866	9,263	15,974
	Nonresidential	15,149	28,810	26,149
	Residential - All	2,483	4,426	109,828

Note: Confusion matrices showing the classification of pixels in the final, composite product for Hyderabad in 2019 compared to the actual classification in the ground-truth rasters for all pixels with known LULC. The first confusion matrix shows the full, six-category comparison, whereas the second shows the same pixels in a "reduced" three-category format, where all residential categories are treated as one. The colors indicate how categories are aggregated to create this reduced version. The overall accuracy in the six-category context was 59.1 percent; in the reduced three-category typology, 73.2 percent.
Source: Authors.

captured more meaningful, broadly applicable patterns. Note that the areal and roads models are completely separate; their results are presented and should be interpreted accordingly.

3.1.1 Category aggregation

One element of the performance data presented below requires additional explanation. The areal models are six-category classifiers, and each data table displays the F2 scores corresponding to those categories. However, each table also displays a seventh score, "Residential – All." This represents the aggregation of the four types of residential LULC into a single category, as illustrated in Table 4.

This aggregation facilitated comparison between the new areal models and older benchmarks from previous research phases, where our areal models were limited to three categories: open space, nonresidential, and (aggregated) residential. We were able to confirm, for example, that expanding from three- to six-category prediction did not in itself reduce performance, in terms of distinguishing between these three basic types.[41]

Predicting across the full typology represented a major, new challenge. Generally speaking, our areal models were good at identifying residential LULC. However, many of the features used by the Atlas methodology to distinguish between residential subtypes—sidewalks, streetlights, fire hydrants, and other small-scale physical infrastructure—are essentially invisible to sensors operating at 10- or 20-meter resolution. In that context, even limited success differentiating between residential subtypes is an encouraging early return; furthermore, even when the actual subtype is in doubt, we can be fairly confident in the more general assertion that a given area represents residential LULC of some sort.

3.1.2 Direct model output: Withheld samples from training images

The India models were trained using several images from each of 17 Indian cities listed in Table 1, while the Mexico models used two images from each of 9 Mexican cities. Our first assessment of model quality used the reserved validation samples from those same images. We anticipated fairly strong performance here, as the models should be tuned to precisely these images (though not the

Table 5 | **Validation F2 Scores for India Models: Training Imagery**

MODEL	SIX CATEGORY AREAL									BINARY ROADS
CITY	F2 SCORE BY LULC CATEGORY							AVERAGE F2 SCORE		F2 SCORE
	OPEN SPACE	NONRESIDEN-TIAL	RESIDENTIAL - ATOMISTIC	RESIDENTIAL - INFORMAL	RESIDENTIAL - FORMAL	RESIDENTIAL - PROJECT	RESIDENTIAL - ALL	SIX-CATEGORY (FULL)	THREE-CATEGORY (REDUCED)	ROADS
Ahmed-abad	0.6891	0.4357	0.0582	0.4923	0.3282	0.4072	0.8565	0.4018	0.6604	0.5851
Belgaum	0.7296	0.4816	0.2339	0.5476	0.5175	X	0.8523	0.5021	0.6878	0.5835
Hindupur	0.8298	0.2240	X	0.7644	X	0.0080	0.7458	0.4565	0.5999	0.3924
Hyderabad	0.6359	0.4004	0.0308	0.4643	0.6605	X	0.8937	0.4384	0.6433	0.5326
Jaipur	0.6952	0.4283	0.0224	0.6396	0.2807	0.0242	0.8609	0.3484	0.6615	0.5809
Kanpur	0.7068	0.4487	0.0315	0.7429	0.2649	0.1295	0.8445	0.3874	0.6667	0.4928
Kolkata	0.7148	0.4971	0.8284	0.0119	0.3219	0.1168	0.8915	0.4151	0.7011	0.4285
Malegaon	0.6444	0.5127	0.0910	0.5772	0.4301	X	0.8129	0.4511	0.6567	0.5068
Mumbai	0.7235	0.6098	0.4983	0.0196	0.3066	0.2879	0.7512	0.4076	0.6948	0.5913
Parbhani	0.8085	0.2145	0.0275	0.8033	0.0239	X	0.8685	0.3755	0.6305	0.5025
Pune	0.6416	0.4674	0.2471	0.4201	0.6514	0.2312	0.7996	0.4431	0.6362	0.5880
Singrauli	0.7561	0.5752	0.0791	0.1368	0.0948	0.6929	0.7522	0.3892	0.6945	0.4306
Sitapur	0.7435	0.3525	0.0418	0.6858	X	0.0919	0.8709	0.3831	0.6557	0.4106
Vijayawada	0.8106	0.4442	0.2990	0.3820	0.5345	X	0.8249	0.4941	0.6932	0.4882
Average (by city)	0.7164	0.4501	0.1997	0.4440	0.3569	0.2249	0.8337	0.4121	0.6667	0.5048
Standard deviation	0.0584	0.1025	0.2460	0.2695	0.2033	0.2076	0.0481	0.0396	0.0243	0.0615

Note: Scores generated via direct application of trained models to reserved validation samples from the training imagery. The final column describes performance of the dedicated, binary roads model, whereas the other data columns represent the multicategory areal model. X indicates that there were no validation samples in that city corresponding to that category, so no score could be calculated.
Source: Authors.

specific sites represented in the validation data). For each model, we scored each image and then took the average score across all images of each city, as shown in Tables 5 and 6. Two aspects of these tables are worth noting. First, the various residential subtypes exhibit highly variable and often quite low scores. In many cases, these are consequences of very small sample sizes. Not every residential subtype is present in every city, and even when present, there may be only a handful of instances out of tens of thousands of samples.

Second, geospatial image "shifting" can impact scoring by causing training data and model output to be misaligned. The performance specification for Sentinel-2 calls for absolute geolocation uncertainty to remain below 20 meters, so a given image may plausibly be offset by several 5-meter pixels in any direction (ESA n.d.d). When superimposed, supposedly congruent scenes occasionally displayed just this effect. This georegistration error can affect our modeling at multiple points in the methodology, but its most apparent impact is with scoring. A model might accurately map the LULC as seen in an image, but that map may be misaligned with the ground-truth data and therefore obtain a low score when the two are compared.[42]

Past experience has indicated that quantitative scoring can be especially misleading when assessing the quality of roads maps. Because the road features are often only one or two pixels wide, even a small geospatial discrepancy between ground-truth and image can reduce or eliminate the overlap between where the roads are "supposed to be"

Table 6 | **Validation F2 Scores for Mexico Models: Training Imagery**

MODEL	SIX-CATEGORY AREAL									BINARY ROADS
CITY	F2 SCORE BY LULC CATEGORY							AVERAGE F2 SCORE		F2 SCORE
	OPEN SPACE	NONRESIDEN-TIAL	RESIDENTIAL - ATOMISTIC	RESIDENTIAL - INFORMAL	RESIDENTIAL - FORMAL	RESIDENTIAL - PROJECT	RESIDENTIAL - ALL	SIX-CATEGORY (FULL)	THREE-CATEGORY (REDUCED)	ROADS
Culiacán	0.8363	0.5724	X	0.5560	0.6929	0.0044	0.8950	0.5324	0.7679	0.6332
Guadalajara	0.8086	0.6791	X	0.6314	0.4948	0.0936	0.7403	0.5415	0.7427	0.6491
Mérida	0.8006	0.5474	X	0.1253	0.2433	0.0717	0.5722	0.3577	0.6401	0.4101
Mexico City	0.7188	0.6445	0.2254	0.4959	0.7094	0.0635	0.7809	0.4763	0.7147	0.6671
Monterrey	0.8160	0.7002	0.0085	0.1999	0.5472	0.1648	0.5150	0.4061	0.6771	0.5430
León	0.8251	0.3996	0.0985	0.5231	0.7545	X	0.8488	0.5202	0.6911	0.5206
Reynosa	0.7888	0.5823	0.0024	0.5193	0.2345	0.6778	0.7832	0.4675	0.7181	0.5466
Tijuana	0.6157	0.6335	0.0629	0.2735	0.5096	0.3158	0.6139	0.4018	0.6210	0.5308
Tuxtla	0.8488	0.5754	0.0966	0.2726	0.3184	0.0410	0.6047	0.3588	0.6763	0.4833
Average (by city)	0.7843	0.5927	0.0824	0.3997	0.5005	0.1791	0.7060	0.4514	0.6943	0.5538
Standard deviation	0.0692	0.0840	0.0743	0.1714	0.1879	0.2090	0.1256	0.0684	0.0443	0.0786

Note: Scores generated via direct application of trained models to reserved validation samples from the training imagery. The final column describes performance of the dedicated, binary roads model, whereas the other data columns represent the multicategory areal model. X indicates that there were no validation samples in that city corresponding to that category, so no score could be calculated.
Source: Authors.

and where they appear in the imagery. This may explain an observed weak relationship between quantitative and qualitative assessments. Manual inspection and comparison with satellite imagery allows for an immediate and usually convincing, albeit subjective, impression of the quality of the roads mapping. A human observer simply knows what a road network typically looks like. This contrasts with the areal LULC products, whose quality is difficult to assess by direct inspection.

3.1.3 Direct model output: Novel imagery

When applying the India model across a large area like the entire country, our system automatically selected the five "best" (i.e., least cloudy)[43] satellite captures for a given area and time period and then applied the model to each, as described in section 2.6.3 "Training and validation tranches." These images were all completely foreign to the model, although some imagery of course covered the Indian cities used for training data. To estimate the quality of this model output, we combined the LULC model outputs corresponding to the least cloudy image for each geospatial tile in order to create a single, comprehensive LULC mosaic of each city for which ground-truth was available. Due to the scarcity of training data, we did not reserve for validation any entire cities, only sets of locales. We then evaluated that mosaic relative to the ground-truth as before, with results shown in Table 7. Because the roads model was not applied at scale in India, there are no corresponding scores. Similarly, our Mexico models were not applied at scale in the same fashion, and so are not represented here.

3.1.4 Final composite product

As described in section 2.8.4, "Post-processing," the final data product for India is a mosaic of LULC tiles, each of which represents a composite of multiple LULC maps generated directly by the model. Scoring was conducted as before, with results shown in Table 8.

An even more granular inspection of model performance can be facilitated with a "complete" confusion matrix, representing the classification of all validation locations from all cities in India—some 1.5 million pixels—in the

Table 7 | **Validation F2 Scores for India Model – Novel Imagery**

MODEL	SIX-CATEGORY AREAL								
CITY	F2 SCORE BY LULC CATEGORY							AVERAGE F2 SCORE	
	OPEN SPACE	NONRESIDEN-TIAL	RESIDENTIAL - ATOMISTIC	RESIDENTIAL - INFORMAL	RESIDENTIAL - FORMAL	RESIDENTIAL - PROJECT	RESIDENTIAL - ALL	SIX-CATEGORY (FULL)	THREE-CATEGORY (REDUCED)
Ahmedabad	0.6503	0.5387	0.0325	0.1393	0.4627	0.4085	0.7589	0.3720	0.6493
Belgaum	0.6217	0.5247	0.1688	0.3502	0.4199	0.0000	0.5954	0.3476	0.5806
Hindupur	0.8325	0.1208	X	0.5700	0.0000	0.0079	0.5459	0.3062	0.4997
Hyderabad	0.6427	0.3512	0.0605	0.4183	0.6351	0.0000	0.8776	0.3513	0.6238
Jaipur	0.7042	0.4827	0.0058	0.6179	0.3245	0.0064	0.8269	0.3569	0.6713
Kanpur	0.7182	0.3967	0.0071	0.7384	0.0197	X	0.8349	0.3760	0.6499
Kolkata	0.6690	0.4978	0.3316	0.0030	0.1645	0.0073	0.5640	0.2789	0.5769
Malegaon	0.5677	0.3820	0.0153	0.4344	0.2695	X	0.8153	0.3338	0.5883
Mumbai	0.6924	0.6087	0.5208	0.0117	0.3038	0.2693	0.7451	0.4011	0.6821
Parbhani	0.6801	0.2197	0.0294	0.6056	0.1325	0.0000	0.8899	0.2779	0.5966
Pune	0.6281	0.4155	0.2803	0.1741	0.5882	0.2390	0.8033	0.3875	0.6156
Singrauli	0.7634	0.6176	0.0394	0.1089	0.0416	0.5912	0.7199	0.3604	0.7003
Sitapur	0.7882	0.3358	0.2537	0.0063	0.0000	X	0.2578	0.2768	0.4606
Vijayawada	0.7760	0.3725	0.2825	0.4380	0.1199	0.0053	0.6298	0.3324	0.5928
Coimbatore	0.7936	0.5356	0.0277	0.5870	0.1312	0.0301	0.6394	0.3509	0.6562
Jalna	0.7535	0.4433	0.0308	0.5416	0.3057	0.0109	0.7590	0.3476	0.6519
Kozhikode	0.6075	0.4557	0.0796	0.0047	0.0000	0.0000	0.2843	0.1912	0.4492
Average	0.6936	0.4255	0.1660	0.3233	0.2363	0.1398	0.7240	0.3394	0.6144
Standard deviation	0.0639	0.1130	0.1671	0.2597	0.2059	0.2009	0.1761	0.0424	0.0626

Note: Scores generated via direct application of trained models to automatically selected, novel imagery from the months after the 2019 monsoon. Coimbatore, Jalna, and Kozhikode are included here, but the necessary data were not available at the time Table 5 was generated. X indicates that there were no validation samples in that city corresponding to that category, so no score could be calculated.
Source: Authors.

final, composite LULC map. We also include recall and precision for each category. We again present results in the context of the full, six-category typology, as in the composite map (Table 9), as well as the three-category results with all residential types aggregated (Table 10). As elsewhere, these scores do not compensate for image shifting (discussed in section 3.1.2, "Direct model output: Withheld samples from training images"), although the compositing process may ameliorate this issue.

3.1.5 India scores comparison

Table 11 facilitates a direct comparison of these different stages of production. The scores based on the full six-category typology are quite noisy, as discussed in section 3.1.2, "Direct model output: Withheld samples from training images," but the aggregated three-category scores display some discernible trends. Comparing scores for the final composite product with those for the novel imagery demonstrates the value of the compositing stage of the workflow. The columns for novel imagery represent the LULC maps based on automatically selected satellite images—and in fact the least cloudy, "best" image for each location—which comprise the inputs to the compositing process. The final, composite map performs better than the individual maps it is derived from.

According to the scores, the composite map also performs nearly as well as the direct application of the model to validation data from the actual images used for train-

Table 8 | **Validation F2 Scores for India Model: Composite Map**

MODEL	SIX-CATEGORY AREAL								
CITY	F2 SCORE BY LULC CATEGORY							AVERAGE F2 SCORE	
	OPEN SPACE	NONRESIDEN-TIAL	RESIDENTIAL - ATOMISTIC	RESIDENTIAL - INFORMAL	RESIDENTIAL - FORMAL	RESIDENTIAL - PROJECT	RESIDENTIAL - ALL	SIX-CATEGORY (FULL)	THREE-CATEGORY (REDUCED)
Ahmedabad	0.6817	0.5098	0.0379	0.4978	0.3160	0.4377	0.8492	0.4135	0.6802
Belgaum	0.6917	0.5201	0.3642	0.4592	0.3752	0.0000	0.6684	0.4017	0.6267
Hindupur	0.7975	0.2932	X	0.8092	X	X	0.7992	0.6333	0.6300
Hyderabad	0.6172	0.3758	0.0198	0.4243	0.5876	0.0000	0.8948	0.3374	0.6293
Jaipur	0.6507	0.5277	0.0047	0.6484	0.3583	0.0074	0.8635	0.3662	0.6806
Kanpur	0.7165	0.5588	0.0028	0.7370	0.0267	X	0.8422	0.4083	0.7058
Kolkata	0.7574	0.5819	0.6779	0.0027	0.3672	0.0484	0.7596	0.4059	0.6996
Malegaon	0.5549	0.4083	0.0042	0.5483	0.4535	X	0.8289	0.3938	0.5974
Mumbai	0.7247	0.6627	0.3796	0.0497	0.2490	0.2097	0.5672	0.3792	0.6515
Parbhani	0.7411	0.2452	0.0136	0.7111	0.1932	0.0000	0.8911	0.3174	0.6258
Pune	0.6408	0.4018	0.2402	0.3182	0.6523	0.1439	0.8093	0.3995	0.6173
Singrauli	0.7774	0.6110	0.1105	0.1592	0.1301	0.6947	0.7612	0.4138	0.7165
Sitapur	0.7847	0.4370	0.0892	0.5264	0.0000	0.0251	0.7319	0.3104	0.6512
Vijayawada	0.7985	0.5543	0.3756	0.2179	0.5540	X	0.8074	0.5001	0.7200
Coimbatore	0.8265	0.5215	0.0254	0.6195	0.1517	0.0920	0.6661	0.3728	0.6714
Jalna	0.7704	0.4916	0.0288	0.6332	0.1510	X	0.7461	0.4150	0.6694
Kozhikode	0.6102	0.4787	0.1064	0.0181	0.0000	0.0000	0.2903	0.2022	0.4597
Average (by city)	0.7122	0.4897	0.1485	0.4010	0.2768	0.1221	0.7471	0.3730	0.6497
Standard deviation	0.0800	0.1051	0.1941	0.2588	0.2133	0.2024	0.1530	0.0658	0.0642
Overall (all pixels)	0.7286	0.5373	0.4306	0.5280	0.4167	0.2771	0.7438	0.4864	0.6699

Note: Scores generated via direct application of trained models to automatically selected, novel imagery from the months after the 2019 monsoon. Coimbatore, Jalna, and Kozhikode are included here, but the necessary data were not available at the time Table 8 was generated. X indicates that there were no validation samples in that city corresponding to that category, so no score could be calculated.
Source: Authors.

ing. Just as the ability of the model to generalize to areas beyond the training data is a critical element of success, so too is its capacity to accurately classify LULC across different periods of time—that is, different images. Our results indicate that our workflow allows us to generate LULC maps derived from novel imagery comparable in quality to those based on known, "already seen" imagery.

3.2 Maps

The data product presented here is the set of final maps produced by compositing model outputs. Parts of some of these maps are shown below. Because of their relatively high resolution, most information is lost in citywide images—only a few large-scale features and patterns remain visible. However, full-resolution rasters are available on the Resource Watch platform, which allows a user to view the maps at a wide range of scales.

Table 9 | **Confusion Matrix: Indian Validation Locations in Composite Map—Full Six-Category**

LULC CATEGORY MODEL		MODEL						CATEGORICAL STATISTICS	
		OPEN SPACE	NONRESI-DENTIAL	RESIDENTIAL - ATOMISTIC	RESIDENTIAL - INFORMAL	RESIDENTIAL - FORMAL	RESIDENTIAL - PROJECT	RECALL	PRECISION
Ground-truth	Open Space	62,866	9,263	131	9,640	4,927	1,276	0.726886704	0.735304029
	Nonresidential	15,149	28,810	1,535	8,668	14,818	1,128	0.544158273	0.511494735
	Residential - Atomistic	355	1,149	182	4877	3,231	22	0.398183611	0.638974163
	Residential - Informal	1,000	1,339	131	16,280	12,147	122	0.557261262	0.436305794
	Residential - Formal	786	1,864	357	17,486	53,374	25	0.42227663	0.396007091
	Residential - Project	342	74	15	94	393	1,092	0.299753249	0.212650251

Note: Overall accuracy (the rate of correct classification) of the six-category LULC map was 56.4 percent.
Source: Authors.

Table 10 | **Confusion Matrix: Indian Validation Locations in Composite Map—Aggregated Three-Category**

LULC CATEGORY MODEL		MODEL			CATEGORICAL STATISTICS	
		OPEN SPACE	NONRESIDENTIAL	RESIDENTIAL - ALL	RECALL	PRECISION
Ground-truth	Open Space	62,866	9,263	15,974	0.726886704	0.735304029
	Nonresidential	15,149	28,810	26,149	0.544158273	0.511494735
	Residential - All	2,483	4,426	109,828	0.739894583	0.759939182

Note: Overall accuracy (the rate of correct classification) of the LULC map, with residential types aggregated, was 68.9 percent.
Source: Authors.

- India
  - Six-category areal product (2016 and 2019): https://resourcewatch.org/data/explore/cit049-Urban-Land-Use-India_1
- Mexico[44]
  - Six-category areal product (2018): https://resourcewatch.org/data/explore/cit047-Urban-Land-Use-Mexico
  - Binary roads product (2018): https://resourcewatch.org/data/explore/cit048-Urban-Land-Use-Mexico-Roads

The raster assets underlying these maps are also available for inspection on Google Earth Engine.

3.2.1 India

The maps of India cover all urban areas. This means, for one thing, that the cities used for training are mapped in their entirety. For example, we can view an areal LULC map of Hyderabad in 2016 (Figure 10).

Viewing a small section of Hyderabad allows for a better subjective assessment of the quality of the results. For more detail, we can view a semitransparent version of the areal LULC map on top of high-resolution satellite imagery (Figure 11).

The maps of urban India extend far beyond the small set of training cities from the Atlas, allowing us to examine LULC in a city like Bangalore, for which we had no ground-truth. Furthermore, because we cre-

Table 11 | **Comparison of Average F2 Scores at Different Stages of Mapping, India**

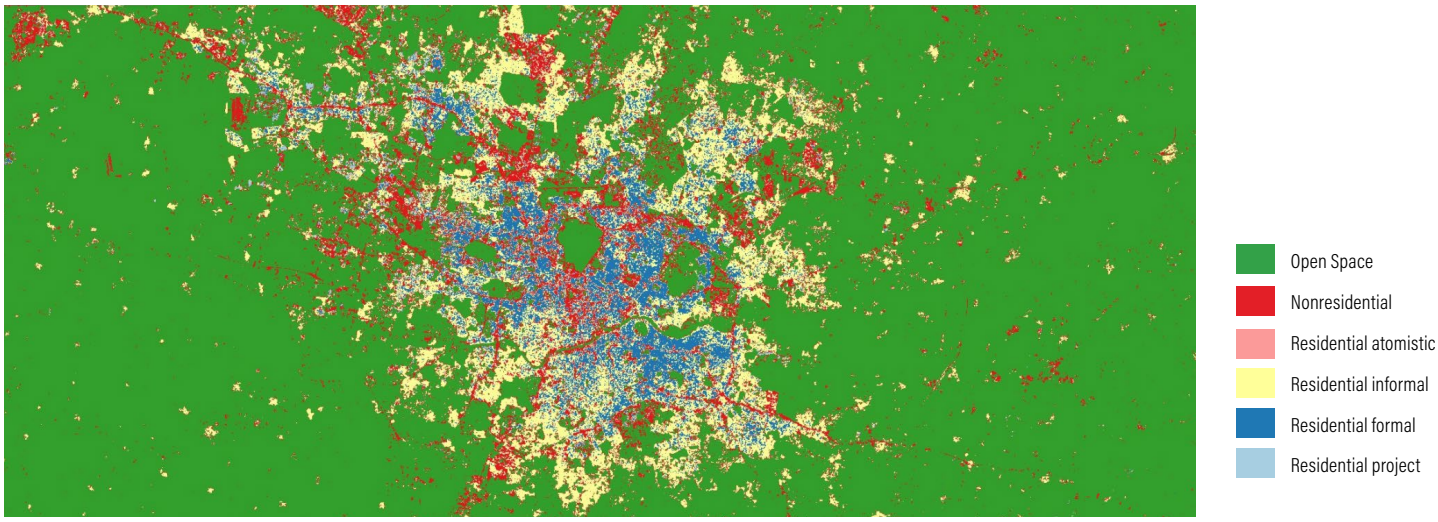
CITY	AVERAGE F2 SCORE - SIX-CATEGORY			AVERAGE F2 SCORE - THREE-CATEGORY		
	TRAINING IMAGES	NOVEL IMAGERY	COMPOSITE MAP	TRAINING IMAGES	NOVEL IMAGERY	COMPOSITE MAP
Ahmedabad	0.4018	0.3720	0.4135	0.6604	0.6493	0.6802
Belgaum	0.5021	0.3476	0.4017	0.6878	0.5806	0.6267
Hindupur	0.4565	0.3062	0.6333	0.5999	0.4997	0.6300
Hyderabad	0.4384	0.3513	0.3374	0.6433	0.6238	0.6293
Jaipur	0.3484	0.3569	0.3662	0.6615	0.6713	0.6806
Kanpur	0.3874	0.3760	0.4083	0.6667	0.6499	0.7058
Kolkata	0.4151	0.2789	0.4059	0.7011	0.5769	0.6996
Malegaon	0.4511	0.3338	0.3938	0.6567	0.5883	0.5974
Mumbai	0.4076	0.4011	0.3792	0.6948	0.6821	0.6515
Parbhani	0.3755	0.2779	0.3174	0.6305	0.5966	0.6258
Pune	0.4431	0.3875	0.3995	0.6362	0.6156	0.6173
Singrauli	0.3892	0.3604	0.4138	0.6945	0.7003	0.7165
Sitapur	0.3831	0.2768	0.3104	0.6557	0.4606	0.6512
Vijayawada	0.4941	0.3324	0.5001	0.6932	0.5928	0.7200
Coimbatore	n/a	0.3509	0.3728	n/a	0.6562	0.6714
Jalna	n/a	0.3476	0.4150	n/a	0.6519	0.6694
Kozhikode	n/a	0.1912	0.2022	n/a	0.4492	0.4597
Average (limited)	0.4121	0.3394	0.3847	0.6667	0.6144	0.6632
Standard deviation	0.0396	0.0424	0.0506	0.0243	0.0626	0.0413
Average (all)	0.4121	0.3302	0.3730	0.6667	0.6082	0.6497
Standard deviation	0.0396	0.0540	0.0658	0.0243	0.0722	0.0642

Note: Data for the cities of Coimbatore, Jalna, and Kozhikode were not available until after the first set of scores was generated. Accordingly, we calculated two averages: the first using only the 14 cities used across all quantitative assessments, and the second using all 17 cities where available.
Source: Authors.

ated comprehensive LULC maps for 2016 and 2019, we can compare the urban form of an area at two points in time (Figure 12).

Certain areas of application also demonstrate the limitations of the current areal model for India. Such insights had to be drawn from inspection of the final LULC maps, as we have no ground-truth outside of the cities listed in Table 1. In the northwest, for example, the final LULC output is implausibly skewed toward nonresidential LULC (Figure 13). Over the course of the project, we have repeatedly observed this type of aberration in areas with high topographic variation, such as foothills and mountains. This may ultimately be a matter of training data: our ground-truth is heavily concentrated in fully urbanized or heavily populated areas, and cities rarely feature canyons and gorges, or the heavy shadowing that comes with steep terrain. In a similar vein, the northwestern

Figure 10 | Hyderabad, Areal LULC, 2016



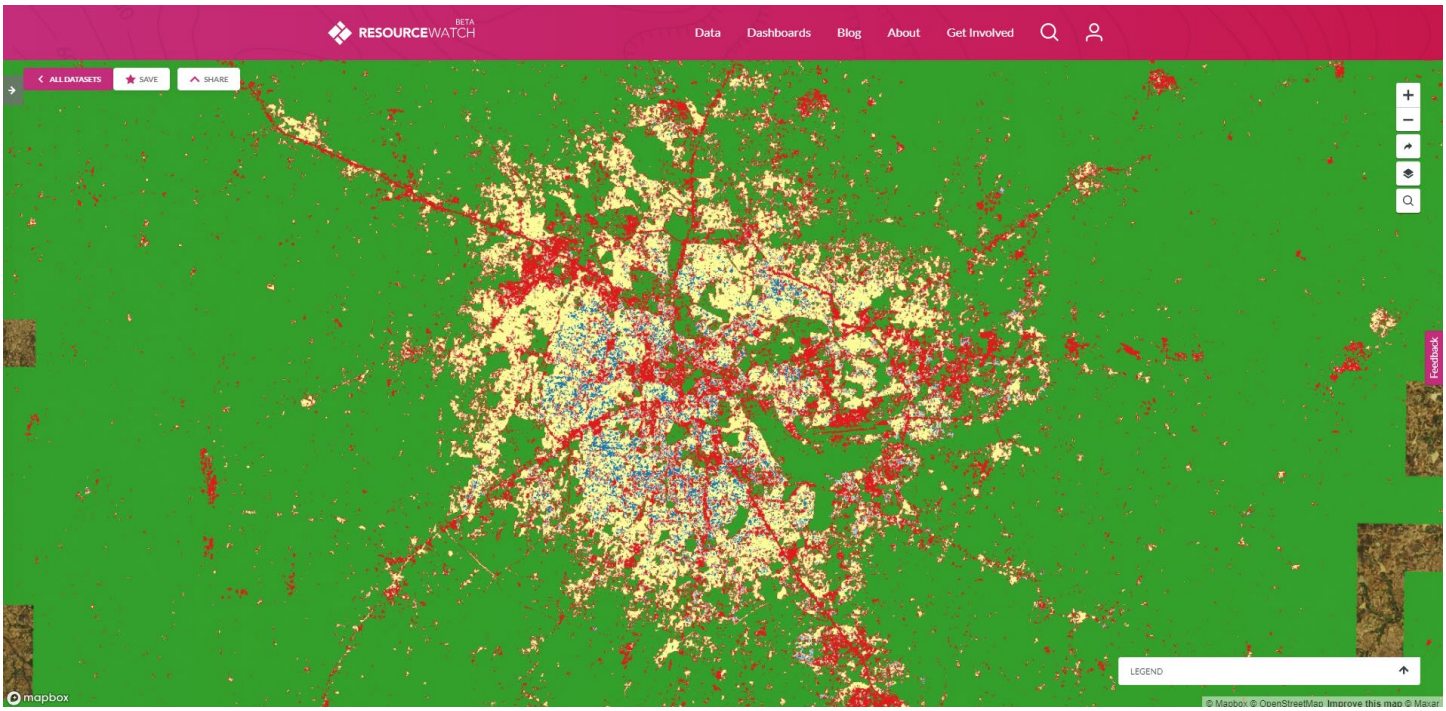
Sources: Authors.

Figure 11 | Hyderabad Neighborhood, Areal LULC with High-Resolution Satellite Imagery

Sources: Satellite imagery accessed via Mapbox; LULC map by authors.

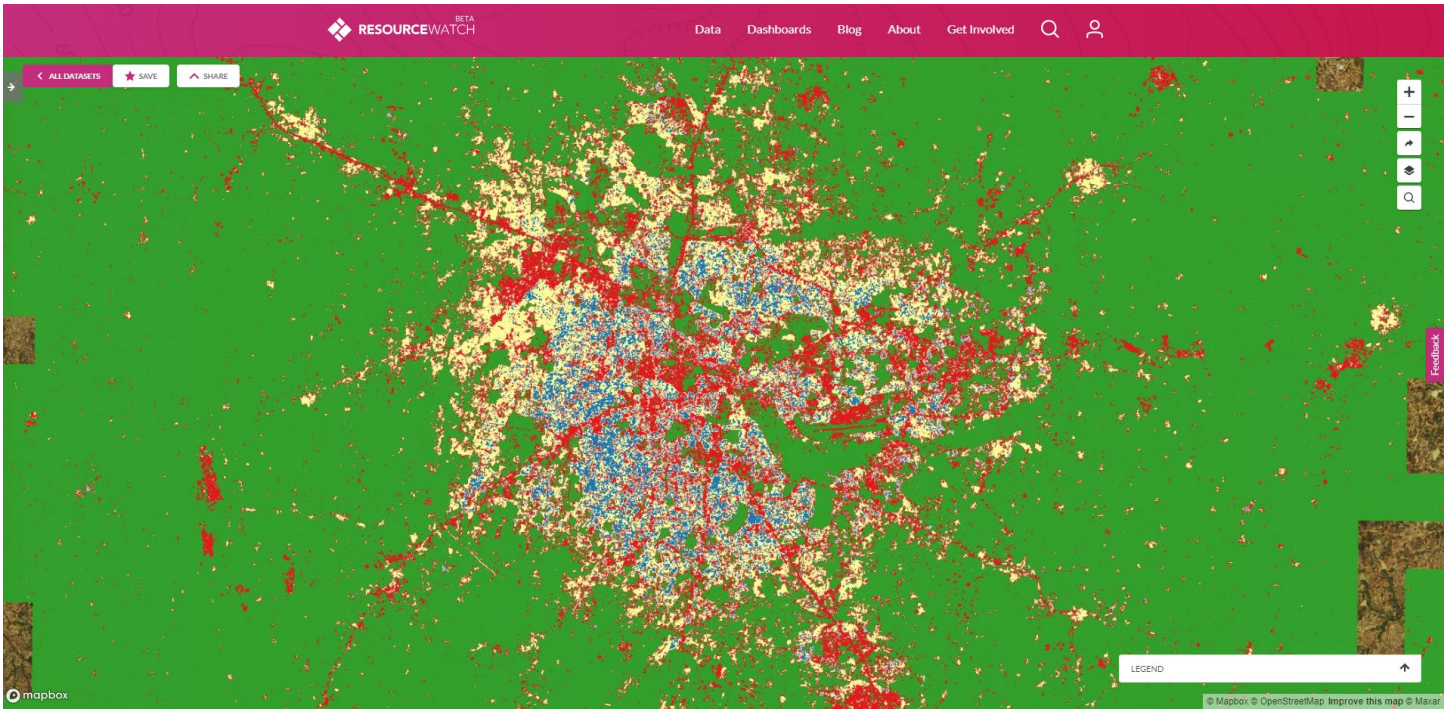
Figure 12 | Bangalore, Areal LULC, 2016 and 2019

Bangalore, 2016



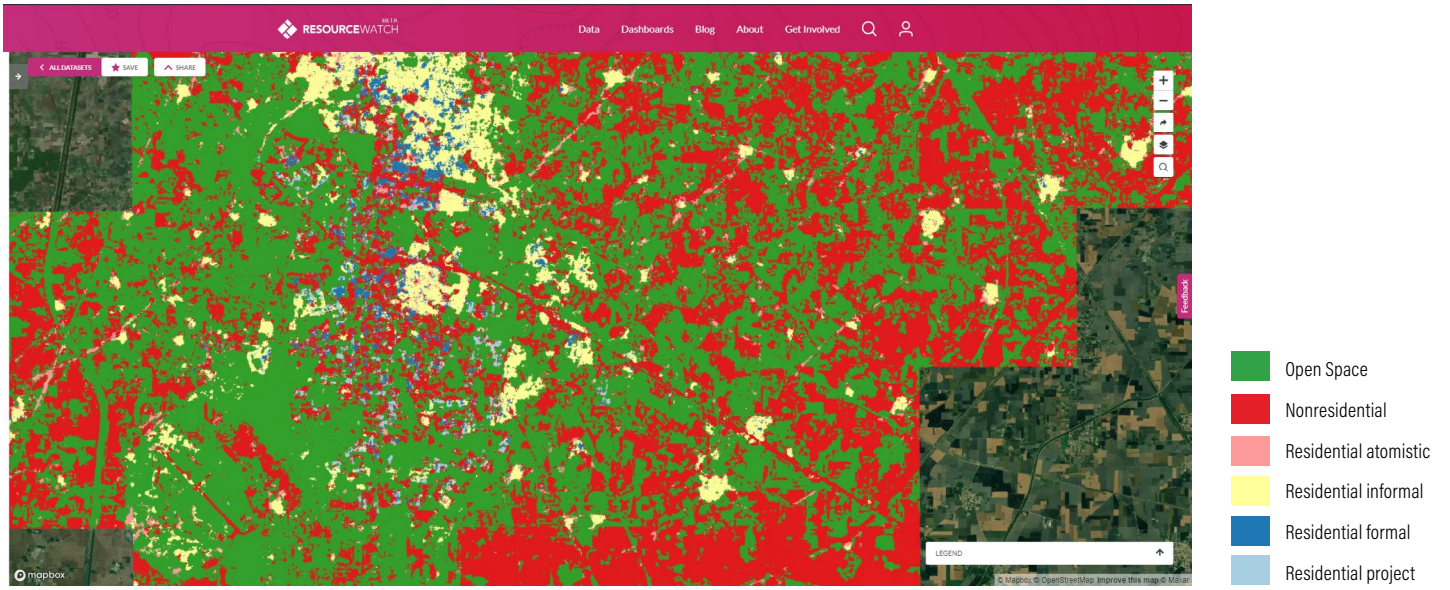
Open Space
Nonresidential
Residential atomistic
Residential informal
Residential formal
Residential project

Bangalore, 2019



Note: LULC in Bangalore, as characterized by composited model output, in 2016 (top) and 2019 (bottom). These maps indicate, for example, growth in the main nonresidential clusters to the west of the city, a new cluster to the east, and expansion of informal housing in the satellite district to the northeast of the city core.
Source: Authors.

Figure 13 | **Modeled LULC near Firozpur in Northwestern India, 2019**



Note: The data product shows a typical variety of LULC types within the city itself, but the nonresidential type (incorrectly) dominates in the hinterland, which in reality is primarily farmland and natural land cover. Satellite imagery is visible where LULC was not classified.
Source: Satellite imagery accessed via Mapbox; LULC map by authors.

territory shown in Figure 13 lies within an ecoregion—in essence, a type of ecosystem and geography—that does not overlap with any of the Indian Atlas cities from which we drew our ground-truth (Dinerstein et al. 2017).[45] This distinct landscape may simply be too novel for our model, particularly outside of truly urban areas where natural land cover is more prominent.

3.2.2 Mexico

In Mexico, the roads model consistently generated highly coherent road networks, particularly in urban cores. Areal LULC and roads are stored separately; the roads product can be overlaid atop the areal product for a full characterization of the urban/periurban environment, as in Figure 14.

Road network coherence generally held at both the city and neighborhood scale. Note that the displayed road networks are direct model outputs, without any morphological post-processing to "clean" or "complete" roads. Figure 15 shows modeled roads in Monterrey at the citywide and neighborhood scales.

Areal maps of Mexican cities revealed another large-scale pattern: a concentration of formal[46] housing around the city core, giving way to increasingly informal types at the periphery, particularly in areas of recent or active expansion (Figures 16 and 17). These areas are often peppered with residential projects, as cities presumably seek to meet the burgeoning demand for housing demonstrated by informal residential districts.

A closer look reveals details of the intermixing of LULC types at the urban periphery. Figure 18 shows a district in the southeast of Mexico City (seen in the lower-right portion of Figure 17). Moving away from the city center, formal residential LULC gives way to informal, save for some well-delineated patches of formal LULC, interspersed with housing projects. On the outermost fringes are undefined smatterings of "atomistic" housing, the least formal category of residential LULC. Nonresidential (i.e., commercial or industrial) LULC runs throughout all of these, although it is noticeably less present in areas of atomistic housing.

3.3 Applying the Methodology Elsewhere

The methodology presented here is explicitly designed for application in other geographies. The Atlas includes cities from around the world. By using ground-truth data from a given area in conjunction with corresponding imagery, new models can be created and new LULC maps generated. This is demonstrated by our use of identical techniques in India and Mexico: nothing changed other than the pools of training data used.

Figure 14 | **Culiacán, Modeled Areal LULC and Roads**

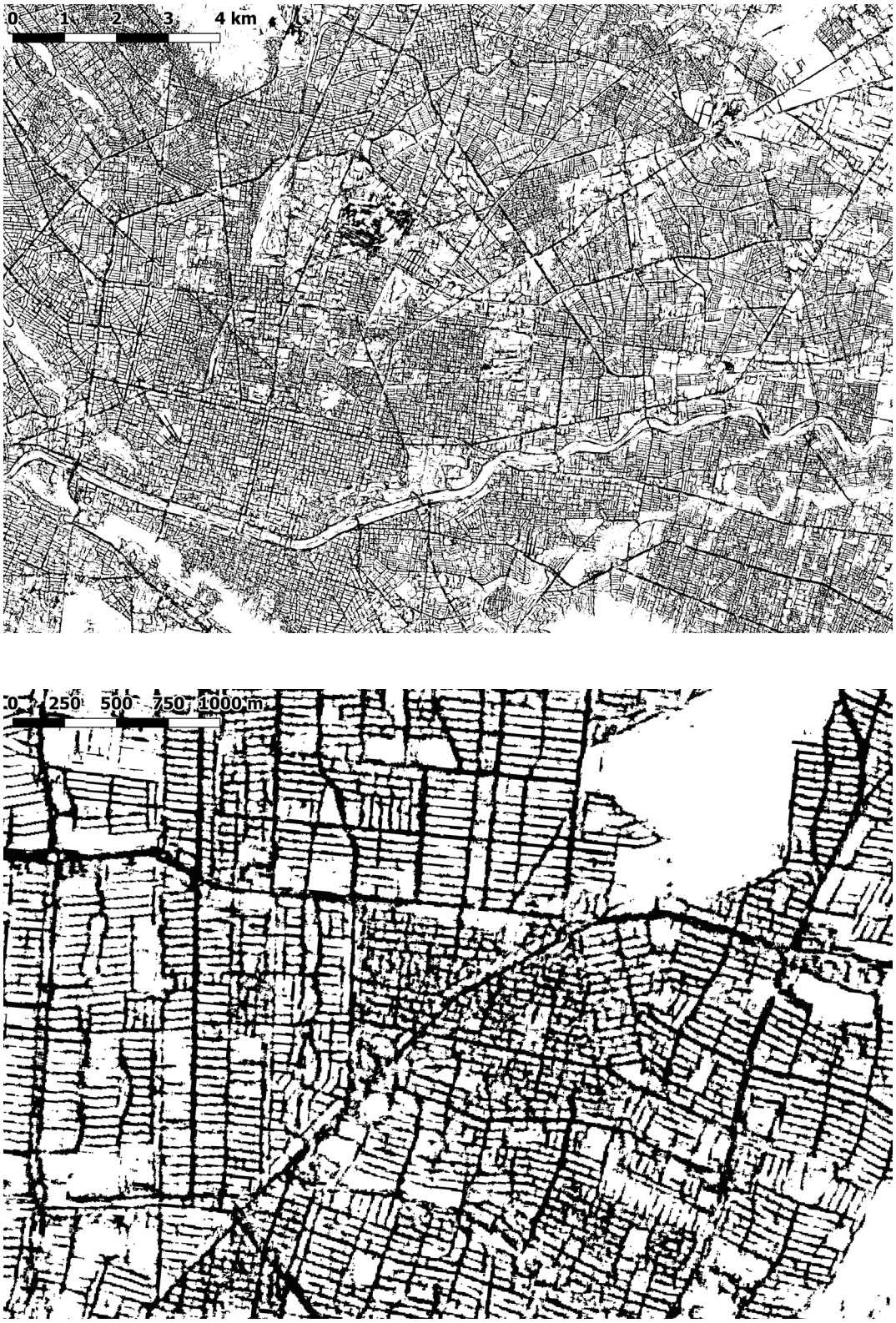
Notes: Six-category areal model output overlaid with roads model output, both derived from the same images of Culiacán from late 2018.
Source: Authors.

For mapping India, we used all available Indian ground-truth in the Atlas. Similarly, our maps of Mexican cities were generated by models trained using all available ground-truth from cities in Mexico. A training data set need not align with national boundaries. While not published here, we used the exact same process to create a regional areal model using ground-truth from all eight countries assigned to the Atlas's "South and Central Asia" region.[47] When applied to Indian cities, this regional model achieved quantitative performance comparable to the India-only model presented in this paper. This success may be attributed at least in part to the prevalence of Indian ground-truth even in the larger regional pool—17 of the 32 regional cities lie in India —but nonetheless suggests the viability of such an approach. We did not systematically evaluate the performance of the regional model outside of India.

For those looking to replicate this approach in other geographies, the primary challenge is to construct an appropriate pool of training data. Some cities and countries are more similar than others, in terms of the built environment. Proximity is often an indicator of similarity, but not always: see how the Atlas itself groups European cities with those in Japan. With that said, national and regional groupings have shown themselves to be generally effective, but users may nevertheless choose to exclude certain "anomalous" cities or include additional ones, as judgment and testing suggest.

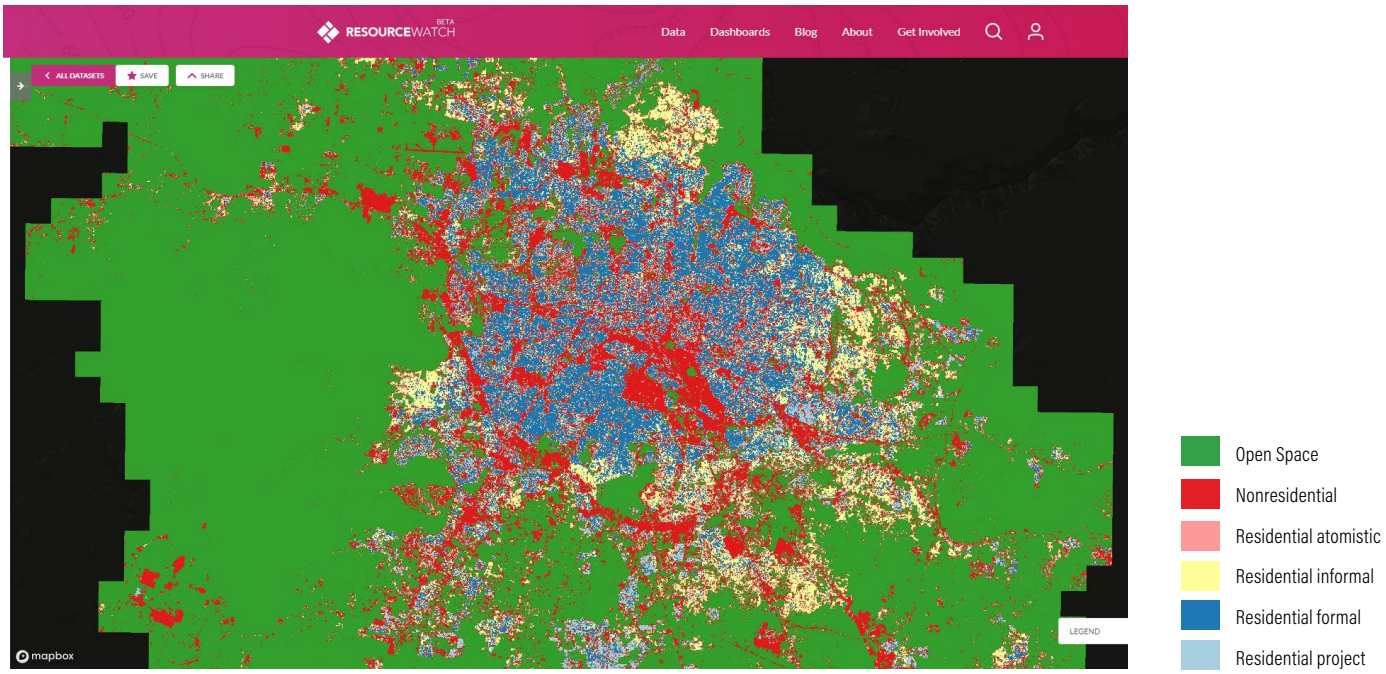
The goal is to have the training data reflect the full area of application to the greatest extent possible, so that the resulting model is able to make meaningful predictions across the entirety of the target geography. In the case of India in the work presented here, we used all available Indian training data, yielding models generally success-

Figure 15 | **Monterrey, Modeled Citywide and Neighborhood Roads, 2018**



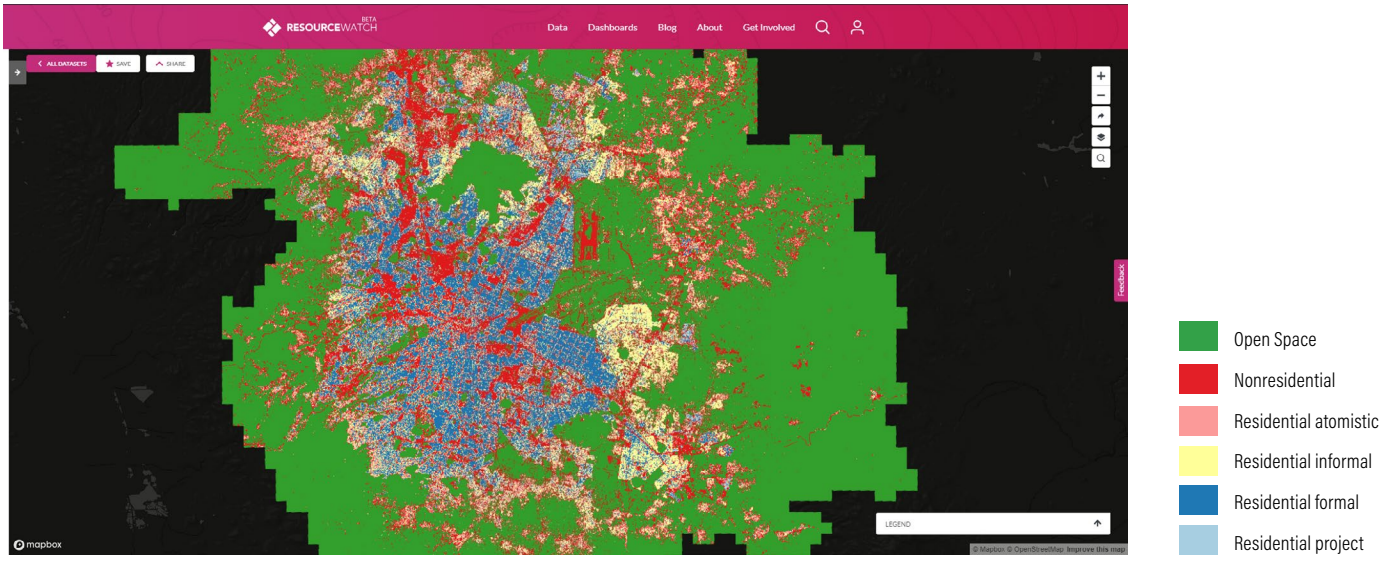
Notes: Modeled road networks of Monterrey, based on imagery from late 2018. The contours of the road network are visible at both the city scale (top) and the district or neighborhood level (bottom).
Source: Authors.

Figure 16 | Guadalajara, Modeled Areal LULC, 2018



Notes: Modeled six-category areal LULC in Guadalajara, based on imagery from late 2018. Light yellow areas on the outskirts of the city core represent informal housing; pink fringes in those areas represent even more unstructured "atomistic" housing. Dark denotes unmapped space beyond the study area.
Source: Authors.

Figure 17 | Mexico City, Modeled Areal LULC, 2018



Notes: Modeled six-category areal LULC in Mexico City, based on imagery from late 2018. Light yellow areas to the east of the city core represent informal housing, whereas the pink areas ringing the city represent even more unstructured "atomistic" housing. Dark denotes unmapped space beyond the study area.
Source: Authors.

Figure 18 | **Detail of Mexico City, Satellite Imagery and Modeled Areal LULC, 2018**

Notes: Modeled six-category areal LULC in Mexico City, based on imagery from late 2018, as shown on Resource Watch, WRI's open data platform. Light yellow areas to the east of the city core represent informal housing, whereas the pink areas ringing the city represent even more unstructured "atomistic" housing.
Source: Authors.

ful across the country. But, as noted, the model was less effective in the distinctive terrain of the northwest, for which we had no comparable ground-truth. In the case of Mexico, we relied on local experts to advise us which cities should be added to our pool of ground-truth to have a fair representation of all different "types" of urban areas in the country. Generally speaking, however, the availability of suitable ground-truth remains a constraint.

Once the cities for inclusion in training data have been selected, the process of model creation can proceed just as described in this document. The Atlas ground-truth must be processed and rasterized, and local imagery must be acquired. After storing the relevant files and creating a catalog of samples, a model can be initialized and trained. Obviously, a user must arrange for their own computing environment, as well as access to appropriate satellite imagery.[48] Application of a trained model, whether for evaluating performance or generating maps, can also simply follow the routines described here and accessible through the project Github repository.

3.4 Quality and Utility of Results

Ultimately, assessing the quality and utility of these models and their outputs is challenging. As discussed above, the quantitative metrics are likely impacted by georegistration inaccuracies. Another factor not discussed in this paper is the difficulty of ascribing land use within an extended typology: how often would even human analysts disagree as to the classification of a given parcel of land? Our quantitative scoring methods do not admit the complexities of ambiguity.

Inspection of the maps is also of limited help. As noted, the quality of a map of roads is relatively easy to assess, because the structure of a road network is well-known and relatively standardized: roads are usually a uniform width, roads are often gridlike, and so on. But the areal LULC is almost impossible to assess without comparing classifications against high-resolution imagery, and even then, the true class may be indeterminate, at least without considerable training. And besides that, it is impractical to perform such assessments at any scale—otherwise, there would be no need for automated LULC classification in the first place.

Lastly, it is difficult to characterize the utility of these LULC maps in the absence of comparable alternative data sets. There are no LULC data sets covering a wide range of urban areas, let alone at five-meter resolution or using a similar typology. In the future, we may attempt to leverage cadastral maps for this purpose, but beyond the questions of their existence, availability, and digitization, there is also no guarantee that such maps would be up-to-date and accurate. In the end, these urban LULC maps represent new, promising information whose value can be best determined by use.

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
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
Governance	Compare planning or zoning data to maps showing current land use	▪ Governments can assess adherence to zoning or other land use regulations ▪ Academics can evaluate impact of zoning on subsequent development
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 actors 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

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

TOPIC	USE CASE	AUDIENCE
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
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
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 can work 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, "City Selection."

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

3. Additional information at Maxar: https://www.maxar.com/products/building-footprints.

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 U.S. 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 programme 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. All code and scripts are available at the project Github repository: https://github.com/wri/UrbanLandUse/releases/tag/v0.2.0.

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

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

11. Detailed information and full specifications for the Tesla K80 GPU can be found on the NVIDIA website: https://www.nvidia.com/en-gb/data-center/tesla-k80/.

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

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

14. 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.

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

16. 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.

17. 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.

18. The Atlas website includes visualizations of fully classified locales superimposed on high-resolution satellite imagery. This may be the best way to become familiar with what structures or neighborhoods were typically assigned to which LULC category. This content can be found within the "Blocks and Plots" section of each city's profile; for example, see the page for Hyderabad: http://www.atlasofurbanexpansion.org/cities/view/Hyderabad#blocks_and_plots.

19. 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: https://github.com/wri/UrbanLandUse/tree/v0.2.0/aue-preprocessing.

20. 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.c). As such we did not investigate the relative merits of these different products.

21. To "upsample" imagery is to 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.

22. 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.

23. Images used for the construction of training data were manually selected (after some automated filtering). This differs from application of the model in India, where satellite imagery was automatically selected for ingestion and mapping. Manual imagery selection was motivated by a desire to ensure that model performance was not limited by image quality, particularly as we made our first investigations into widely applicable (i.e., nationally or more broadly) LULC classifiers. Furthermore, this selection need be performed only once per area with training data. Nevertheless, based on the success of automatic image selection for model application, in the future we will likely seek to similarly automate image selection for training data creation.

24. 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.

25. Each 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 sematic segmentation-type model, for example, where the output is effectively a grid of classifications congruent to the input look window.

26. This means that the machine-learning algorithm, seeking to establish patterns relating local appearance to LULC, could "see" at least eight pixels—40 meters—in any direction from the central location (i.e., the pixel the algorithm sought to classify).

27. 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 5 m/pixel, despite having mixed native resolutions.

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

29. 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.

30. According to its website, "pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool, built on top of the Python programming language" https://pandas.pydata.org/.

31. The areal models could predict any of the six nonroad LULC categories, and so had six output nodes. The roads models were binary, predicting only "road" or "not road," and thus had only two output nodes. This discrepancy was the only structural difference.

32. The full mechanics of the fairly simple loss function are defined in the project code at https://github.com/wri/UrbanLandUse/blob/0da4c3560caa0f625148172634fa8d141637fe12/utils/util_training.py#L65.

33. Information about the "Tasks API" is available at https://docs.descarteslabs.com/guides/tasks.html.

34. The 2016 map represents the window starting November 1, 2016; the 2019 map, November 1, 2019.

35. According to the native Level-1C product cloud layer.

36. Where there were only two votes and they disagreed, the category corresponding to the higher raw, numerical model output was given precedence.

37. Confusion matrices are a very common way of comparing model predictions to actual, correct values. For a description, illustration, and links to additional material, visit: https://scikit-learn.org/stable/modules/model_evaluation.html#confusion-matrix.

38. 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 at https://scikit-learn.org/stable/modules/model_evaluation.html#precision-recall-f-measure-metrics. The scores presented in the results tables represent the macro-average F-score.

39. For definitions of precision and recall, as well as a discussion of their relationship and related modeling tradeoffs, see: https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html.

40. The most straightforward metric for performance is 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.

41. These findings are not presented here, but an interested reader can easily compare between the results presented here and those in the previous technical note (Kerins et al. 2020).

42. The ground-truth was derived from very high-resolution imagery with much smaller geolocation error. For example, the Airbus Pleiades 1-A satellite claims an image location accuracy of three meters without ground control points, or up to one meter with them (Satellite Imaging Corporation n.d.). We thus expect that any georegistration-based errors stem primarily from our mapping imagery, rather than the ground-truth.

43. As elsewhere, this was determined using the cloud mask layer native to the Level-1C product.

44. A map of the Mexican roads data product laid over the six-category areal product is available here: https://bit.ly/3iSQ3T1.

45. The referenced data set, RESOLVE Ecoregions by Biome, can be accessed and viewed on the Resource Watch platform: https://resourcewatch.org/data/explore/bio042-Ecoregion-by-Biome. The territory in northwestern India shown in Figure 13 is described in this data set as ecoregion "Aravalli west thorn scrub forests," an assignation not shared by any of the areas used for ground-truth.

46. Recall that this terminology refers exclusively to observable physical features, and has no connection to the legal status of a housing unit or neighborhood.

47. The included countries are Afghanistan, Bangladesh, India, Iran, Kazakhstan, Nepal, Pakistan, and Uzbekistan. All cities belonging to each region can be viewed on the website of the Atlas: http://www.atlasofurbanexpansion.org/cities.

48. Sentinel-2 data is free to download; see ESA's Sentinel Online at https://sentinel.esa.int/web/sentinel/sentinel-data-access. Landsat imagery is also freely and directly available from NASA, as described at https://landsat.gsfc.nasa.gov/data/where-to-get-data/. However, given the volume of data required for some applications, it may behoove users to employ a programmatic approach, perhaps mediated by a dedicated platform.

REFERENCES

Angel, Shlomo, Alejandro M. Biel, Jason Parent, Patrick Lamson-Hall, and Nicolás Galarza Sánchez. 2016a. *Areas and Densities*, vol. 1 of *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. *Blocks and Roads*, vol. 2 of *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-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.

Chollet, François. 2017. "Xception: Deep Learning with Depthwise Separable Convolutions." *ArXiv:1610.02357 [Cs]*, April. http://arxiv.org/abs/1610.02357.

Corbane, Christina, Aneta Florczyk, Martino Pesaresi, Panagiotis Politis, and Vasileios Syrris. 2018. "GHS Built-Up Grid, Derived from Landsat, Multitemporal (1975-1990-2000-2014), R2018A." European Commission, Joint Research Centre. doi: 10.2905/jrc-ghsl-10007.

Dinerstein, Eric, David Olson, Anup Joshi, Carly Vynne, Neil D. Burgess, Eric Wikramanayake, Nathan Hahn, et al. 2017. "An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm." *BioScience* 67 (6): 534–45. doi:10.1093/biosci/bix014. Accessed through Resource Watch (April 20, 2020). https://resourcewatch.org/.

EEA (European Environment Agency). 2017. "Copernicus Land Monitoring Service—Local Component: Urban Atlas." https://land.copernicus.eu/user-corner/publications/ua-flyer/.

EEA. 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.

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

ESA. n.d.b. "Level-1C Product Formatting - Sentinel-2 MSI - Technical Guide - Sentinel Online." https://sentinel.esa.int/web/sentinel/technical-guides/sentinel-2-msi/level-1c/product-formatting.

ESA. n.d.c. "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. n.d.d. "Performance—Sentinel-2 MSI Technical Guide—Sentinel Online." https://earth.esa.int/web/sentinel/technical-guides/sentinel-2-msi/performance.

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): 1164–76. doi:10.1109/JSTARS.2012.2190383.

Kerins, Peter, Emily Nilson, Eric Mackres, Taufiq Rashid, Brook Guzder-Williams, and Steven Brumby. 2020. "Spatial Characterization of Urban Land Use through Machine Learning." Technical Note. Washington, DC: World Resources Institute. https://www.wri.org/publication/spatial-characterization-urban-land-use.

Mahdianpari, Masoud, Bahram Salehi, Mohammed Rezaee, Fariba Mohammedimanesh, and Yun Zhang. 2018. "Very Deep Convolutional Neural Networks for Complex Land Cover Mapping Using Multispectral Remote Sensing Imagery." *Remote Sensing* 10 (7): 1119. doi: 10.3390/rs10071119.

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.

Satellite Imaging Corporation. n.d. "Pleiades-1A Satellite Sensor | Satellite Imaging Corp." https://www.satimagingcorp.com/satellite-sensors/pleiades-1/.

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). 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.

USGS. 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.

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 would like to thank the National Geographic Society's Geographic Visualization Lab for both funding and counsel for this research. 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:

- Steven Brumby, Sam Brooks Hyde, Joe Mazzariello, and Fabien Laurier at National Geographic Society
- Patrick Lamson-Hall, Professor Shlomo Angel, Nicolas Galarza Sanchez, and Alejandro Blei, 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 continued special thanks to Jeremy Malczyk, Sam Skillman, and Jay Carlson
- Our WRI Mexico colleagues Jorge Macias, Natalia Garcia, Mauricio Brito, Gorka Zubicaray, Abraham Berumen, and Alejandra Bosch
- Reviewers Leslie Dewan, Thomas Esch, Caitlin Kontgis, Patrick Lamson-Hall, Alejandra Bosch, John Brandt, Mauricio Brito, Samantha Kuzma, Gregory Taff, and Thet Hein Tun

ABOUT THE AUTHORS

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

Brookie Guzder-Williams is the Director of Data Science 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 Analyst for the Research, Data, and Innovation program at the World Resources Institute.

Eric Pietraszkiewicz was a Remote Sensing and Machine Learning Intern 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.