NBER WORKING PAPER SERIES

DETECTING URBAN MARKETS WITH SATELLITE IMAGERY: AN APPLICATION TO INDIA

Kathryn Baragwanath Vogel
Ran Goldblatt
Gordon H. Hanson
Amit K. Khandelwal

Working Paper 24796
http://www.nber.org/papers/w24796

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2018

We acknowledge funding from the World Bank, the International Growth Centre (Project 89448), and the Center for Global Transformation at UC San Diego. We thank the Editor, Gilles Duranton, and two anonymous referees for valuable feedback. Additionally, we thank Somik Lall, Trevor Monroe, Rinku Murgai, Adam Storeygard, Siddharth Sharma, and seminar participants at University of Toronto, Berkeley Haas, 2018 Urban Economics Association Meetings, the Urban and Regional/Spatial Zoom Seminar, MIT, McGill University, and the World Bank for constructive comments. Khandelwal acknowledges support from The Council on Foreign Relations International Affairs Fellowship in International Economics. All errors are our own. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Detecting Urban Markets with Satellite Imagery: An Application to India
Kathryn Baragwanath Vogel, Ran Goldblatt, Gordon H. Hanson, and Amit K. Khandelwal
NBER Working Paper No. 24796
July 2018, Revised May 2019
JEL No. O1,O18,R1

ABSTRACT

We propose a methodology for defining urban markets based on builtup landcover classified from daytime satellite imagery. Compared to markets defined using minimum thresholds for nighttime light intensity, daytime imagery identify an order of magnitude more markets, capture more of India's urban population, are more realistically jagged in shape, and reveal more variation in the spatial distribution of economic activity. We conclude that daytime satellite data are a promising source for the study of urban forms.

Kathryn Baragwanath Vogel
University of California, San Diego
9500 Gilman Dr
La Jolla, CA 92093
kbaragwa@ucsd.edu

Ran Goldblatt
New Light Technologies
1100 H St NW #700
Washington, DC 20001
ran.goldblatt@nltgis.com

Gordon H. Hanson
IR/PS 0519
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0519
and NBER
gohanson@ucsd.edu

Amit K. Khandelwal
Graduate School of Business
Columbia University
Uris Hall 606, 3022 Broadway
New York, NY 10027
and NBER
ak2796@columbia.edu

1 Introduction

Core to the study of economic geography is explaining why cities exist and how their dimensions are determined. It is standard to attribute the existence of cities to the benefits of agglomeration—be they urbanization economies (e.g., Henderson 1974), Marshallian externalities (e.g., Duranton and Puga 2001), or home-market effects (e.g., Fujita et al. 2001). Where cities locate, in turn, is influenced by the availability of key resources, access to transportation routes, and historical accident (e.g., Bleakley and Lin 2012, Henderson et al. 2018). Within cities, the clustering of activity creates gradients in land prices and presents workers with a tradeoff between housing costs and commute times (Duranton and Puga 2015). A rich and vibrant literature studies how the concentrating forces of agglomeration and the dispersing forces of congestion combine to create urban systems (e.g., Duranton and Puga 2004, Desmet and Henderson 2015).

Empirical work on economic geography requires measuring the location and scale of urban activity. A common approach to measurement is to use officially designated administrative units. These may be as large as a metropolitan area (e.g., Duranton and Turner 2012), as small as a town or village (e.g., Eeckhout 2004), or an intermediately sized unit such as a county or a district (e.g., Hanson 2005, Ghani et al. 2014, Donaldson and Hornbeck 2016). Because administrative boundaries are defined according to pre-existing legal jurisdictions, they may be noisy indicators of how cities are actually organized. In influential work, Rozenfeld et al. (2011) construct cities by clustering officially designated towns and villages into larger units based on geographic proximity. This approach only works, however, if official sources measure activity for fine administrative units on a frequent basis. In many countries, such data are available only decadally, if at all.

In this paper, we use remotely sensed data to detect urban markets in India for 2013. A *market* is a set of contiguous, or near contiguous, pixels that contain economic activity according to daytime or nighttime satellite imagery. Our practical approach approximates the conceptual definition of a market in economic geography models: a set of locations that are highly integrated relative to outside locations because of low internal trade costs (e.g., Redding 2016) and (or) low commuting costs (e.g., Duranton 2015). We categorize a pixel as having economic activity if its nighttime light intensity exceeds a given threshold or its spectral properties indicate builtup landcover. Our maintained assumption is that clusters of proximate pixels are integrated through internal trade and commuting links, which we attempt to validate in external data.

Our first source of imagery is nighttime lights from the Defense Meteorological Satellite Program Operational Linescan System, which indicates the presence of economically active agents (Henderson et al. 2012). Following Rozenfeld et al. (2011), we explore buffers that combine contiguous sets of pixels if they lie within a radius of $1km$, $2km$, $4km$ or $8km$. Defining urban land using nightlights requires choosing a minimum threshold of light intensity for contiguous pixels. Harari (2017), for instance, in her analysis of urban sprawl in large Indian cities chooses a digital number (DN) of 35 (on a scale of 0 to 63) to designate urban areas. Our analysis reveals a tradeoff in choosing the minimum light threshold for a market: while a strict threshold only captures major urban agglomerations, lowering the threshold to include smaller cities explodes the size of larger

cities with proximate satellites. This tradeoff is a consequence of the blooming effect of light, which produces cities whose boundaries are too expansive and too smooth relative to actual cities.

We contrast the spatial extent of nightlight-based markets with those formed from high-resolution daytime satellite imagery. These data are available at finer resolutions than nighttime lights data but require further image classification to detect urban land. We explore data on builtup landcover from the MODIS layer constructed by Channan et al. (2014). We also examine two additional daytime imagery layers: the Global Human Settlements Layer (Pesaresi et al. 2015) and a recent layer produced by Goldblatt et al. (2018). We define landcover-based markets using an analogous algorithm that clusters contiguous or near contiguous pixels of builtup landcover.

Our approach has three advantages over conventional methods to measure urban areas using administrative data. First, it is scalable. Because our method is algorithmic and uses publicly available imagery, it scales to detect markets globally and, in principle, over time. It also circumvents the need to reconcile differences across countries and time in how administrative units are defined. Second, and relatedly, our data do not stop at national borders. Markets that straddle national boundaries along transportation routes can be tracked. Third, the spatial resolution is adjustable. By altering the buffer used to aggregate proximate pixels, we can narrow the focus to the rough equivalent of a town center or widen the focus to a metropolitan area. This versatility is helpful for detecting within-metro area heterogeneity, a feature we explore.

To preview our results, the patterns of landcover-based markets are starkly different from those of nightlight-based markets. Using the definition of a market that buffers clusters of contiguous pixels at $1km$, we detect 1,669 and 469 markets using a nightlight threshold of DN33 and DN60, respectively. The DN60 markets accurately capture India's largest 470 cities according to official Census data, which suggests that nightlight-based markets are reliable for tracking activity across India's major urban areas. In contrast, we detect an order of magnitude more markets using MODIS data: 12,953 in total. These markets are smaller, less compact, more closely fit a power law in area size, and capture activity ranging from distinct areas within large metropolises to small towns that are distant from India's major cities. For example, within Delhi's official administrative boundary, we detect 579 distinct $1km$ MODIS markets. More remote, landcover-based markets have an average DN nightlight intensity of just 5, suggesting that we are able to capture many parts of India that lack reliable access to electricity. While we could detect these markets with nightlight data by lowering the light-intensity threshold, this would come at the cost of vastly increasing the area of above-threshold contiguous pixels around India's large cities, which is evident from visual inspection and from statistics on the maximum size of markets at different thresholds. Our results suggest that landcover-based markets are able to capture small cities and towns in India while preserving the spatial distribution of activity of the largest cities.

We perform several validation checks to demonstrate that our markets capture real economic activity. Using shape files for the 2011 India Census, we allocate population across our market boundaries. Collectively, the DN33 and DN60 $1km$ markets contain 23.4% and 14.8% of India's total population and 75.3% and 47.6% of India's urban population, respectively. The MODIS $1km$

markets capture 29.0% and 93.2% of India's total and urban populations, respectively. Market size correlates strongly with population, and the variance in population for smaller sized landcover markets reflects the fact that these markets include both dense areas within major metro areas and less populated peripheral towns. We detect strong correlations between market size and proximity to public infrastructure, such as roads, railway stations and mobile phone towers. Additionally, we find that larger landcover-based markets have higher nighttime light intensity. These correlations are important for addressing a limitation of daytime satellite data. While these data are suitable for measuring the spatial extent of markets, they may not reveal the intensity of economic activity. However, the positive correlations reveal that the extensive margin—which is measured accurately through daytime imagery—correlates well with proxies for the intensive margin. For example, a MODIS market at the $10th$ percentile of the land-area distribution has a nightlight DN of 9.4 compared to 27.2 for a market at the $90th$ percentile of land area.[1] Combining daytime imagery to measure the boundary of markets with nighttime data to measure the intensity of activity is a promising approach to leverage two remotely sensed datasets that are publicly available, have a long time span, and have global coverage.

Finally, we consider the potential to use landcover-based markets to study polycentric cities (Duranton and Puga 2015). The literature has long recognized that cities do not expand smoothly along their margins but through the construction of outlying communities in the form of suburbs, edge cities, or commercial hubs (e.g., Henderson and Mitra 1996, Anas et al. 1998). For example, the Hyderabad metro area, which spans $650km^2$ and contains 6.8 million people, contains Hyderabad and Secunderabad as major poles and substantial satellites in Ghatkesar and Kukatpally. As one zooms in further, many more satellites appear and Hyderabad's full polycentricity is revealed. We examine polycentricity using the larger buffered markets, which we term "super-markets". The average MODIS $8km$ market spans an area of $63.4\,km^2$, but physical structures cover only 23% of this area. On average, these super-markets contain 4.2 distinct $1km$ markets; the elasticity of the number of $1km$ buffered markets with respect to super-market area size is 0.36. This within-market variation may be sufficient to study, for instance, how transportation investments, such as ring roads or metro rail, impact the distribution of economic activity within large cities. To demonstrate this possibility, we construct measures of market access based on Donaldson and Hornbeck (2016) and find that a non-trivial portion of a market's access is determined by other close-by markets that are within the same larger buffered super-market.

The availability of satellite imagery and machine-learning techniques for image classification have led to rapid advances in detecting land use in the remote sensing literature. In efforts to construct urban layers for the world as a whole, Pesaresi et al. (2015) use Landsat imagery to detect urban land for grid cells at a $38m$ resolution, Channan et al. (2014) use MODIS imagery to detect multiple types of land use for grid cells at a $500m$ resolution, and Zhou et al. (2015) use nightlight intensity to detect urban land at a $1km$ resolution.[2] This work typically classifies land

[1] Using the nightlight-GDP elasticity of 0.3 from Henderson et al. (2012) implies that the larger area would have a 56.8% higher GDP.

[2] Alqurashi and Kumar (2013) discuss earlier work in remote sensing to detect land use. Recent papers that detect

use at the pixel level, where the dimensions of the pixels vary according to the source of the satellite imagery. Pixel-level classifications, while important building blocks in urban analysis, are not in and of themselves informative for the study of economic geography. Without aggregating pixels to form larger markets, one cannot test theories of the size distribution of cities, evaluate the impacts of expanding national transportation grids, or identify the consequences of severe weather events, plant closures, or other localized economic shocks.

Our results contribute most directly to the efforts to delineate urban areas that do not rely on administrative boundaries. In addition to Rozenfeld et al. (2011), our paper has antecedents in Eeckhout (2004), who uses U.S. Census Designated Places instead of (much larger) Metropolitan Statistical Areas to re-examine Zipf's law and Gibrat's law; Burchfield et al. (2006), who use contiguous pixels to measure sprawl in the U.S. based on Landsat satellite imagery from 1976-1992; and Harari (2017), who uses nightlights to track urban sprawl in large Indian metropolitan areas. Davis et al. (2018) also use clusters of pixels above nightlight thresholds to construct metro areas in Brazil, China, and India. Recent work by Duranton (2015) proposes an alternative algorithm to construct markets based on commuting patterns for Colombia. de Bellefon et al. (2018) develop a statistical approach to detect urban areas using precise locational data covering 34 million buildings in France. Our contribution to this literature is to develop and compare methods to detect markets solely from remotely sensed data, and in particular daytime imagery.

More broadly, our paper builds on the increasing use of remotely sensed data for economic analysis. Economists have used satellite data on the intensity of light emitted at night to study national and regional economic growth (Henderson et al. 2012, Gennaioli et al. 2013, Pinkovskiy and Sala-i Martin 2016), the political economy of regional development (Gennaioli et al. 2013, Michalopoulos and Papaioannou 2013a, Michalopoulos and Papaioannou 2013b), spatial linkages between cities (Storeygard 2016), and the global distribution of economic activity (Henderson et al. 2018), among a rapidly growing set of topics. Daytime satellite imagery, whose use in economics was pioneered by Burchfield et al. (2006), is available at even higher spatial resolutions, down to $30m$ for data going back to the late 1990s and down to less than $1m$ for imagery from recently launched proprietary satellites. Michaels et al. (2018) use an ensemble of remotely sensed imagery to measure urbanization in Tanzania. For a comprehensive survey of recent work, see Donaldson and Storeygard (2016).

Section 2 presents the method to detect markets from satellite imagery. Section 3 compares nightlight-based markets and landcover-based markets and provides validation checks. Section 4 uses landcover-based markets to evaluate market access. Section 5 concludes.

2 Algorithmic Approach to Detect Markets

We define markets using two sources of remotely sensed data: (1) the intensity of light as captured by nighttime lights data; and (2) classifications for builtup landcover based on daytime

urban land for individual countries include Pandey et al. (2013) on India; Bagan and Yamagata (2015) on Japan; and Zhou et al. (2014), Huang et al. (2015), and Fu et al. (2017) on China. For literature that detects urban land using daytime satellite imagery, see Trianni et al. (2015), Goldblatt et al. (2016), and Goldblatt et al. (2018).

satellite imagery. In this section, we describe the data sources and algorithms used to detect the spatial extent of a market for each data source.

2.1 Detecting Markets from Nightlight Imagery

The US Air Force Defense Meteorological Satellite Program (DMSP) operates satellites that carry light sensors known as the Operational Linescan System (OLS). Originally used to detect the global distribution of clouds and cloud-top temperatures, OLS sensors also detect visible and near-infrared emissions at night from different sources on Earth, such as city lights, auroras, gas flares, and fires. Pixels have a resolution of 30 arc seconds, or approximately $1km \times 1km$. For each pixel, the digital number of calibrated light intensity ranges from 0 to 63, which we refer to as the nightlight value or intensity. Because persistent light emitted at night is often associated with man-made structures, we assume that if the intensity of a pixel exceeds a given threshold, this pixel represents a populated location. Processed DMSP-OLS imagery is publicly available from 1992-2014, and can be analyzed on Google Earth Engine. We process lit pixels using data for 2013. We use the stable light band of sensor F14, which discards ephemeral events, such as fires, but remains sensitive to persistent lighting, including from gas flares or volcanoes. Since India has no active volcanoes or gas flares on land (Elvidge et al. 1999), it is safe to assume that highly lit pixels in India indicate builtup activity.

There are well-known limitations to DMSP-OLS data. These include saturation effects, in which the amplification of light detection to capture low levels of light leads to right censoring in detection in highly-lit areas (e.g., city centers); and blooming effects, in which reflection causes light emitted in one pixel to be detected in nearby pixels, making highly lit areas appear to be larger than they are. Blooming occurs due to several idiosyncratic features of the DMSP-OLS sensor: (1) field of view variation, where the satellite's round field of view morphs into an elliptical and larger shape as it scans east and west of nadir; (2) geolocation errors, whereby the satellite miscalculates a pixel's location, so on each night not only is there a differently sized ellipse, but its centroid is shocked in a random compass direction (Abrahams et al. 2018); and (3) on-board data management, where the 1970s technology on board the satellites causes top-censoring of inputs. The highest possible DN is 63, and because of this saturation, it is often impossible to differentiate between medium-density cities and high-density cities.[3] In our setting, saturation is not an issue because we measure the extent of markets through lower bounds of light intensity. However, blooming is problematic, as we demonstrate below.

Nightlight-Based Markets: *A nightlight-based market is a cluster of contiguous, or near contiguous pixels, with a DN that exceeds a specified threshold.*

[3]Blooming and saturation are less pronounced in data from recently launched satellites. The Visible Infrared Imaging Radiometer Suite (VIIRS), imagery from which is only available since 2012, detects electric light at a higher spatial resolution and at lower distortion than DMSP-OLS. See Elvidge et al. (2017) for a discussion of VIIRS imagery and Shi et al. (2014) for an application of these data to detecting urban areas in China. Henderson et al. (2018) use a radiance-calibrated version of the nightlight data that alleviates the saturation effect (Elvidge et al. 1999) but these data are also available only for a subset of recent years. We use DMSP-OLS imagery in order to create methods for measuring markets that can be extended backward in time.

To operationalize this definition of a market based on nightlight data, three choices are required: (1) the minimum number of pixels that constitute a market; (2) the parameter values that govern "near contiguity"; and (3) the minimum DN to be used. As mentioned, the DMSP-OLS sensor has a $1km$ resolution. We set the minimum number of pixels to form a market at 1 pixel.[4]

To determine the minimum DN thresholds for our market definition, we examine the distribution of DNs across pixels in India for 2013 in Figure A1. Because light is not detected in large expanses of the country—including bodies of water, farmland, deserts, forests, and villages with no electricity—the DN is zero (i.e., no detectable light) for the pixel at the 50^{th} percentile of the distribution. The DN is moderately higher at a value of 5 at the 63^{rd} percentile, and rises sharply as one moves into the upper tail, reaching 17.4 at the 95^{th} percentile, 49 at the 99^{th} percentile, and 60 at the 99.5^{th} percentile; only a tiny fraction of pixels are right censored at the maximum DN of 63.[5] Motivated by these patterns, we set the following alternative DN thresholds for a pixel to be highly lit: 17.4 (95^{th} percentile), 33 (98^{th} percentile), and 60 (99.5^{th} percentile).

We designate as a market a cluster of contiguous highly lit pixels, which may consist of only a single pixel. Many clusters of highly lit pixels lie in close proximity to each other, creating chains of light islands that appear when we map our results. By the strict definition above, we would treat each island, or polygon of pixels, as a separate market, whereas in truth clusters of proximate polygons may share dense commercial and commuting ties (as in the case of U.S. counties that comprise commuting zones; e.g., Tolbert and Sizer 1996). Motivated by the method in Rozenfeld et al. (2011) for agglomerating neighboring administrative units into larger units, we combine any pair of highly lit clusters for which the minimum distance between their boundaries is less than $1km$, $2km$, $4km$, or $8km$.[6] For a given threshold, larger buffers nest smaller buffers: $1km$ markets $\subseteq$ $2km$ markets $\subseteq$ $4km$ markets $\subseteq$ $8km$ markets.

2.2 Detecting Markets from High-Resolution Daytime Imagery

Daytime imagery offers alternative data to detect human activity from space. The major challenge in working with daytime imagery is that one needs a classifier to convert the spectral signature of an image into a categorization of landcover. In recent years, there has been substantial

[4]The threshold pixel choice of 1 may appear to low. As a point of reference, Rozenfeld et al. (2011) use grid cells with $200m$ resolution for Great Britain and FIPS units for the U.S., which range from $100m$ grid cells in Manhattan to $100km$ grid cells in Wyoming. In recent work, de Bellefon et al. (2018) provide a statistical approach to choose thresholds to define urban areas using detailed geocoded data on the location of buildings in France, and detect distinct urban areas as small as $0.04km^2$.

[5]The bunching at 0 and 5 is an artifact of the stable light band of satellite F14, which removes noise and unstable light removal. Cauwels et al. (2014) note that the number of pixels with DN greater than 0 and less than 5 is extremely low; for example, the satellite registers no pixels with a DN equal to 1 in the year 2000. Tuttle et al. (2014) develop a mapping of DNs to wattage by placing portable high-pressure sodium lamps at uninhabited sites in Colorado and New Mexico to check the DN recorded by the F16 and F18 sensors. They find that ninety-three 100-watt incandescent lamps could be detected (DN=1) at both fine ($0.6km$) and coarse ($2.7km$) resolutions. Eight times as many bulbs would saturate (DN=63) the sensor at the fine resolution but not at the coarser resolution.

[6]We view a $0km$ buffer as extreme as it does not account for commuting or trade linkages and therefore do not consider this buffer choice for our analysis. We use the Aggregate Polygons function in ArcGis to cluster the pixels. Online Appendix A explains the procedure to aggregate pixels to markets.

progress in remote sensing to improve the precision of classification algorithms at scale. Use of daytime imagery is also facilitated by cloud-based computing engines, such as Google Earth Engine, which hosts the full library of Landsat, MODIS, Sentinel, and other satellite imagery.

We use the MODIS dataset as our benchmark source of landcover classification from daytime imagery. MODIS uses a supervised machine learning method, which takes advantage of a global database of training sites extracted from high-resolution imagery that contain 36 spectral bands. We use the University of Maryland classification scheme version MCD12Q1 V006, which has a resolution of $500m$ (Friedl and Sulla-Menashe 2015). We use data from 2013 and take the take the Urban and Builtup pixels (classification 13) to indicate builtup landcover. MODIS is publicly available on Google Earth Engine and widely used in the remote sensing literature (e.g., Huang et al. 2016, Mertes et al. 2015, Guo et al. 2015).[7]

We also examine two other landcover datasets as a robustness check against MODIS. The Global Human Settlements Layer (GHSL, Pesaresi et al. 2015) combines satellite data from Global Land Surveys datasets (GLS1975, GLS1990, GLS2000), Landsat 8, and other sources——including Open Street Maps, WorldPop and MODIS——to determine builtup pixels at a $38m$ spatial resolution.[8] We use their "Builtup Confidence Grid", which aggregates builtup data in 2014 and classifies pixels as builtup if the confidence of being builtup is greater than 50%. GHSL contains landcover maps from an earlier period but has less frequent temporal variation than MODIS. Although publicly available, the GHSL is difficult to access, uses data beyond raw satellite bands, and is less widely used. The third map of builtup landcover for India in 2013 is created using the methodology in Goldblatt et al. (2018). This layer, to which we refer as MIX, uses DMSP-OLS nightlight data as quasi-ground truth and daytime satellite imagery as inputs to train a classifier for builtup land cover in India. Appendix B summarizes their method for producing this layer.

Our motivation for using multiple layers of builtup landcover comes from rapid advances in remote sensing for classifying land use from satellite imagery. The accuracy with which existing layers detect changes in urban landcover, rather than just cross-sectional features, is a subject of on-going research (e.g., see Mertes et al. 2015, Song et al. 2016). We anticipate more advances will be made in land-use classification in the near future, such that none of the existing layers may become the standard source for builtup landcover. In light of this uncertainty, we use three different layers, which allows us to assess the strengths and weaknesses of alternative approaches to detecting urban activity. Our method would easily extend to new layers of builtup landcover.

Using the three layers that classify builtup landcover—MODIS, GHSL, and MIX—we adopt

[7]MODIS (MCD12Q1 V006) classifies global land cover types at yearly intervals from 2001 to 2016. There are six versions of MODIS. The most recent version, Collection 6, improves over previous versions by implementing a hierarchical classification approach, using a RandomForest classifier instead of a C4.5 decision tree, increasing the number of sites in the training data by 47% and updating sites that have changed their land use (about 31% of sites), improving the feature set that now includes phenology metrics, and using Markov chain stabilization. Additionally, we found that this version was slightly better at capturing urban land cover changes in time than its previous version. See the MCD12Q1 V006 user manual for details.

[8]The USGS Landsat 7 satellite, launched in 1999, contains seven spectral bands at a spatial resolution of $30m$ and a temporal frequency of 16 days. Landsat 8, launched in 2013, contains nine spectral bands with a spatial resolution of $30m$ at a temporal frequency of 16 days.

the following definition for markets for daytime satellite imagery.

Landcover-Based Markets: *A landcover-based market is a cluster of contiguous or near contiguous pixels whose spectral features in daytime satellite imagery indicate that the majority of their land area consists of builtup landcover.*

For MODIS markets, we impose a minimum number of pixels for a market to be 1 ($0.25km^2$). For GHSL and MIX, the minimum number of pixels is set to 40 ($0.04km^2$ and $0.03km^2$, respectively). Choosing a minimum pixel size of 1 for GHSL and MIX would be extreme given the granularity of these data (and would be computationally cumbersome); the choice of 40 leverages the granularity of the data to detect small clusters of pixels while not creating markets so small that they would rarely display well-defined internal trade or self-contained commuting patterns. Clusters of builtup pixels are aggregated in a manner analogous to that described above (e.g., if two clusters of MODIS pixels are separated by, say, $1.5km$ of non-builtup pixels, they would form two distinct markets under the $1km$ buffer and a single market under the $2km$ buffer).

2.3 Visual Inspection of Market Definitions

To obtain a visual sense of the shape of urban markets identified by daytime versus nightlight data sources under the four buffers, we plot the markets detected around three cities of different sizes: Delhi (19 million, 2011 Census population), Ahmedabad (5.5 million), and Ajmer (0.5 million) in Figures 1 to 3. We overlay road networks from OpenStreetMaps in 2018 to provide a sense of how transportation networks may influence the shape of markets. Panels (a) to (d), in the first row, display results for MODIS-based markets, while panels (e) to (t), in the second through fifth rows, display results for nightlight-based markets. We include nightlight markets formed using a DN threshold of 10 to understand better the consequences of varying light intensity thresholds but we do not analyze DN10 markets in subsequent sections.

Consider first nightlight-based markets. Together, we have 16 alternative nightlight-based market definitions. The maps illustrate how changing the DN threshold and buffer sizes affects market shape. At a DN of 10 (fifth row), Delhi is an immense blob that swallows cities across three states in India, including Meerut (1.3 million, in Uttar Pradesh), Rohtak (0.4 million, in Haryana) and Bhiwadi (0.1 million, in Rajasthan). The blob itself is $26km^2$, which is close to the size of the U.S. state of Iowa. At a higher DN of 17.4 (fourth row), Delhi takes the shape of a more conventional urban market, but again swallows the city of Meerut (1.3 million), which is $75km$ northeast of central Delhi. At a DN of 60 (second row), by contrast, Meerut appears as a separate market from Delhi. But this threshold fails to detect the small city of Hapur (0.2 million). Moreover, the satellite cities of Gurgaon (0.9 million) and Noida (0.6 million), two vibrant areas of economic activity in Delhi, are fused together with central Delhi to form one large market. Figure 2 for Ahmedabad shows a similar pattern: a high threshold separates the main city from its largest satellite (Nadiad, 0.2 million), but fails to detect many smaller cities; lowering the threshold causes the size of Ahmedabad to explode. Figure 3 shows the smaller city of Ajmer in the state of Rajasthan. The

road leading out of Ajmer towards the Northeast is part of the Golden Quadrilateral. At lower DN thresholds, activity appears to coalesce along the artery. This is problematic as these lights are likely capturing lights along the road rather than stable clusters of economic activity.

To consider landcover-based markets, examine the top rows of Figures 1 to 3, which show markets using the MODIS layer. In stark contrast to the nightlight-based definition in the bottom four rows, landcover-based markets are jagged in shape and display large variation in the spatial density of economic activity. Also, landcover-based markets show that within the outer envelope of the market area there are substantial numbers of white pixel islands, indicating areas that are not builtup. Whereas the blooming effect creates the perception that inside market boundaries all pixels contain light-emitting structures, higher-resolution imagery indicates that cities contain many clusters of pixels that have not been builtup (e.g., undeveloped land, water, and parks). For example, the Yamuna river in Northeast Delhi is visible in the landcover-based figures but masked through the blooming of lights in the nightlight-based markets. The presence of undeveloped pixels within cities in the top row and absence in the lower rows (which are especially apparent in Figures 1 and 2 for the larger cities of Delhi and Ahmedabad and would appear for the smaller city of Ajmer were we to zoom in) indicates that nightlight-based markets tend to make urbanization inside market boundaries appear to be overly smooth. Notice also that within Delhi, we observe many distinct neighborhoods that are fused together in nightlight-based markets. At higher distance buffers, the small distinct markets within cities fuse together while remote towns remain visible.

Visual inspection illustrates the tradeoff in varying the DN threshold to detect markets using nightlights. A strict DN threshold captures the most economically developed urban centers of India. But this threshold misses smaller cities and towns. In attempting to capture these towns through a lower DN threshold, the large cities mushroom in size and swallow neighboring satellite cities. Lower thresholds also start to capture activity along roads which are likely emitted by street lights and (or) the blooming effects from towns. Landcover-based markets detected through high-resolution daytime imagery are not subject to this tradeoff. We observe distinct pockets of activity within cities and detect smaller towns located at the periphery; increasing the buffer fuses together markets within cities while preserving the shape of the smaller cities. Statistics reported in the next section reinforce the descriptive results from this visual inspection.

3 Market Characteristics and Validation

This section explores the characteristics of nightlight- and landcover-based markets based solely on the properties of the satellite data. We then validate that these markets do indeed capture economic activity, by incorporating data from the Indian Census and open source platforms.

3.1 Market Characteristics

We document the following market characteristics. First, while nightlight-based markets capture the largest cities in India, daytime imagery detect an order of magnitude more markets that, on

average, are much smaller in size, are less compact, and have lower nightlight intensities. Second, landcover-based markets capture remote pockets of economic activity, as well as sub-centers within larger urban metropolises. Third, the distribution of landcover-based markets follow a power law that more closely matches Zipf's law than the distribution of nightlight-based markets.

3.1.1 Market Shape

Harari (2017) finds that the geometry of Indian cities affects economic outcomes. Her analysis uses a novel geography-based identification strategy that predicts the compactness of cities, where compactness is measured by how close a city's shape resembles a perfect circle. She determines the extent of cities using a procedure analogous to our nightlight-based markets, and finds that less compact Indian cities have higher commuting costs and lower economic welfare for residents. As shown above, visual inspection suggests that nightlights will produce boundaries that are overly smooth relative to the jagged boundaries of landcover-based markets. If shape determines the welfare of residents, as her study finds, measuring it accurately is important. Her primary measure of urban shape is the *disconnection index*, based on Angel et al. (2010), which is the average distance between all pairs of interior points within a market. In the absence of actual commuting data, the index serves as a proxy for the average commute length within a market.

Figure 4 plots the disconnection index, measured in kilometers, for DN33, DN60 and MODIS markets against market size.[9] For nightlight markets, the disconnection index does not increase with area size. This suggests that the shape of nightlight markets does not fundamentally change with the overall expanse of a market. Increasing market size, by definition, will increase the bilateral distances between some interior points. But the finding that the overall index does not change implies that the market is including builtup pixels in close proximity with other builtup pixels. Thus, the overall compactness appears to be invariant to total market land area. In contrast, the shape of MODIS markets changes starkly with overall market size. As the market land area increases, the disconnection index increases linearly, which reveals that larger MODIS markets are much less compact compared to both smaller MODIS markets and to all nightlight-based markets. For example, the disconnection index of a $100km^2$ MODIS market, buffered at $1km$, is $6.1km$ compared to $3.4km$ and $0.8km$ for DN33 and DN60 markets, respectively. The figure also reveals that disconnectedness increases more sharply for higher buffered markets. These patterns reinforce the visual perception that landcover markets are more jagged and irregular, and therefore more disconnected, than nightlight markets.

[9] The computational burden of computing the disconnection index is very high since it is an average of all bilateral pixels within a market. We therefore only compute this index for the two nightlight-based markets and for MODIS markets. For the same reason, we do not compute the index for $8km$-buffered markets. Harari (2017) normalizes her disconnection index by the average distance between points in a circle that has equivalent area of a given market. We report the disconnection index without normalization, as it is more straightforward to interpret and instead report how the index changes with market size. Additionally, whereas nightlight-based markets consist largely of continuous expanses of lit pixels, MODIS markets contain many undeveloped areas within their outer envelope. It is thus instructive to compare average distances between points within a market without normalizing, since the normalization factor for, say a $4km$ buffer, would be vastly different for MODIS and nightlight markets.

3.1.2 Number of markets

We next explore the number of markets detected through the alternative market definitions. For context, Table A1 reports the official number of enumerations, at various levels of aggregation, according to the 2011 Census. The Census recognizes 6,171 "towns", which are home to India's 377 million urban residents (31% of India's total population).[10] Of the 6,171 towns, 468 are considered "Class 1" cities with more than 100,000 inhabitants; these are the largest cities in India, which collectively contain 22% of India's population. There are 1,847 Class 1, 2 and 3 towns—localities with at least 20,000 inhabitants.

The top panel of Table 1 reports the number of markets detected through nighttime lights. By construction, the number of markets decreases as we raise either the distance buffer for joining pixel clusters or the DN threshold for designating highly lit pixels. At a buffer of $1km$, we observe 3,275 DN17.4 markets, 1,669 DN33 markets, and 469 DN60 markets. The two higher DN thresholds exhibit little variation in the number of markets across buffers. Comparing Table 1 and Table A1, we see that DN17.4 markets at a $1km$ buffer roughly match the number of officially recognized Indian cities and towns with more than 10,000 residents. The DN60 markets accurately capture Class 1 towns, which corroborates the finding in Harari (2017) that nighttime satellite imagery are well-suited for tracking variation in urban form across India's largest cities.

The bottom panel of Table 1 reports the number of markets detected from daytime imagery. While the numbers vary across the three daytime satellite layers, the total number of markets detected is substantially larger than the number of nightlight-based markets. For the MODIS layer, the number of markets ranges from 12,953 at distance buffer of $1km$ to 3,073 at a distance buffer of $8km$. For the GHSL layer, the number of urban markets ranges from 26,202 at distance buffer of $1km$ to 3,861 at a distance buffer of $8km$. The corresponding numbers of markets for the MIX layer are 17,304 and 3,417, respectively.[11] At a distance buffer of $4km$ or less, the total numbers of landcover-based markets are much larger than the number of towns in India with a population of 10,000 inhabitants or greater.

3.1.3 Land area

Column 2 of Table 1 reports the average land area for each market definition. Consider nightlight-based markets, first. For DN17.4, the average size ranges from $48.6km^2$ at a $1km$ buffer to $97.8km^2$ for a distance buffer of $8km$. These values fall, respectively, to $37.0km^2$ and $43.7km^2$ for DN60 markets. These statistics reinforce the tradeoff in choosing a light intensity threshold: lower thresholds detect more markets but the average market size increases. For landcover-based

[10] These towns satisfy one of two criteria: (1) a place with a municipality, corporation, cantonment board, or notified town area committee; or (2) a place that has a minimum of 5,000 inhabitants, at least 75 percent of the male working population engaged in non-agricultural pursuits, and a population density of at least 400 people per km^2.

[11] The GHSL and MIX layers detect more distinct markets in part because the underlying resolution of these data are finer than MODIS. We are unsure of the precise explanation for why the number of GHSL-based markets is so high. Unlike MODIS and MIX, GHSL combines raw daytime spectral bands with MODIS and data from open-sourced platforms, which makes this layer quite different from the other two.

markets, the average market sizes are much smaller. At a $1km$ buffer, MODIS markets are $3.0km^2$, while the average size of GHSL and MIX markets are $1.4km^2$ and $1.9km^2$, respectively. The smaller sizes of landcover markets are a result both of the granularity of the daytime imagery and the exclusion of non-builtup land area (e.g., due to blooming), which we explore in more detail in Section 4. At a $4km$ buffer, the sizes of MODIS, GHSL and MIX landcover-based markets rise to $10.6km^2$, $10.9km^2$, and $12.1km^2$.

To further illustrate the tradeoffs in forming markets with nightlight data, it is useful to compare maximum market sizes. The maximum area of MODIS markets at a $1km$ buffer is $1{,}582km^2$. By contrast, the maximum sizes of nightlight-based markets at a $1km$ buffer changes substantially across the DN17.4, DN33, and DN60 thresholds: from $9{,}977km^2$ for DN17.4 to $4{,}681km^2$ for DN30 and to $2{,}223km^2$ for DN60. To see this further, consider Figures A3 and A4, which plot the distribution of market area and average nightlight values within market boundaries, respectively. Figure A3 reveals that landcover-based markets are able to capture the full range of market sizes. The mode of each distribution effectively reveals the minimum number of pixels used to define a market.[12] Figure A4 illustrates that nightlight-based markets, by construction, are left censored at their respective DN thresholds. Note that because of buffering these markets do capture pixels below their respective thresholds, which is most apparent at the $8km$ buffer. By contrast, at all buffers, landcover-based markets capture pixels that span the entire range of DNs. In particular, these markets capture areas in India with average DNs well below 10.

These comparisons highlight the tradeoff in forming markets from nightlight data. As one lowers the DN threshold to detect smaller markets, the area of larger markets expands dramatically. This tradeoff is not present in the construction of landcover-based markets. Landcover-based markets, because they are not subject to a blooming, span a relatively wide range of land areas and intensities of economic activity (as captured by nightlight intensity per unit of land in these markets).

3.1.4 Power Law of Market Area

Economists have long been interested in the size distribution of cities. The standard approach in the literature is to gather population data using census counts for cities in a particular country and to regress the log of city population on the log city population rank. Zipf's Law holds if the slope of the regression is -1. Testing for Zipf's Law requires confronting the thorny issues of which data sources to use, how to assess the quality of these sources and the accuracy of their implied methods for designating administrative boundaries, and whether to truncate the distribution so as to focus on the properties of the upper tail (Gabaix and Ioannides 2004). The motivation for the algorithmic approach developed by Rozenfeld et al. (2011) is to construct the extent of urban markets without having to rely on seemingly arbitrary boundaries, and then to test for the presence

[12] The right shift of the distribution of land area for nightlight-based markets is most pronounced at a buffer of $1km$, because at this buffer only the high-resolution daytime imagery is able to isolate small urban markets. While the right shift of market-size distributions for the lower-resolution imagery is preserved at higher distance buffers, the relative "peakiness" of the market-size distribution for landcover-based markets diminishes at higher buffers because smaller market areas are joined into larger pixel clusters at these buffers.

of Zipf's Law using cities whose boundaries are justified based on economic fundamentals (i.e., the proximity of their internal clusters of activity). In that paper, they show that the distribution of city land areas approximately obeys Zipf's Law for the US and the UK, and explain that a Zipf's law in area can be rationalized by a model with Cobb-Douglas preferences for goods and housing along with a proportional random growth process.

We examine the emergence of a power law in the distribution of land areas for our market definitions. Following Gabaix and Ibragimov (2011), Figure 6 plots the log of market rank minus 0.5, based on land area, against the log of land area. The figure reveals three patterns. First, landcover-based markets more closely follow the log-linear relationship dictated by a power law. The R^2 of the regressions for landcover-based markets (which range from 0.90 to 0.98) are higher than for nightlight-based markets (which range from 0.84 to 0.91). For MODIS $1km$ markets, the R^2 is 0.96. That is, for landcover markets the entire distribution of market size appears to be Pareto, whereas for nightlight markets the size distribution appears to be Pareto only in the upper tail. Second, the figure also reveals that for nightlight-based markets the shape of the area-rank plot is roughly stable across buffers. This suggests that increasing buffers simply increases the size of markets proportionally, such that the rank-area relationship remains constant. In contrast, the linear slopes of the area-rank plots for landcover-based markets flatten out as the buffer size increases, indicating greater dispersion. Finally, Figure 6 also reveals that for nightlight-based markets, the log-linear relationship breaks down for the largest markets. For landcover-based markets, however, the curve that fits the upper tail markets is close to linearity (as it is in the remainder of the distribution). For the MODIS $1km$ markets, the slope of the line is -0.93.

3.2 Validation

The statistics presented above summarize the extensive margins of urban activity and are based solely on satellite data. A limitation of satellite-inferred markets is that they convey uncertain information on the intensive margin of economic activity. This limitation may be less of a concern with nightlight-based markets, since earlier work demonstrates a strong positive relationship between nightlight intensity and GDP, both in levels and in changes (e.g., Henderson et al. 2012). Daytime satellite imagery, in contrast, provide unknown information on the intensity of economic activity within markets. This is because in the landcover layers the pixels record only whether or not a man-made impervious structure is present. One would need additional information, such as the density and height of structures, to improve the prediction of economic activity based on the underlying spectral signatures of those images.[13]

This subsection matches external datasets to the boundaries of markets to explore correlations between market area and different measures of economic activity. Since we are confident in measuring the land area of a market, the strength of these correlations provides an indication of whether land area is also a reasonable proxy for the intensity of economic activity of cities. We examine cor-

[13] See Jean et al. (2016) for a recent application that predicts micro-spatial poverty headcount for five countries in Africa using nighttime and daytime imagery and Demographic and Health Surveys.

relations between market area, population, nightlight intensity per unit of land, and three granular measures of infrastructure provision——roads, railway stations, and mobile phone towers.

3.2.1 Population

Our first approach to measure economic activity within our market boundaries is to overlay the 2011 India Census to obtain population counts for each market.[14] Census shape files are disaggregated at the town and village level, which have an average area of $16.6km^2$ and $4.8km^2$, respectively (see Table A1). Analogous to Davis et al. (2018), we overlay our markets with the Census towns and villages shape files to spatially match each town to the market it lies inside or overlaps. The population of each town is then assigned to the market it overlaps. If a market overlaps more than one town, the population of all the towns it overlaps is assigned to that market. If a town overlaps more than one market, we divide the population of the town by the number of markets it overlaps, and assign this value to each market it overlaps. This ensures that we are not double counting the population of towns that overlap more than one market.

The third column of Table 1 reports the total population contained in the markets we detect. According to our estimates, the DN60 markets, which as shown above find the Census' Class 1 towns, collectively contain 14.8% of India's population and 47.6% of the urban population. This is lower than the official Class 1 total since DN60 markets identify the core urban area of cities (the DN60 markets are smaller, on average, than the average size of Class 1 towns). The DN33 markets contain 23.4% and 75.3% of India's total and urban populations, respectively.

Compared to these two DN thresholds, landcover markets capture a larger share of India's urban population. The $1km$ MODIS markets contain 29.0% and 93.2% of the total and urban population. Total urban population share rises to 93.8%, 96.7%, and 111.1% for $2km$, $4km$ and $8km$ markets. Thus, we find that landcover markets are able to capture the vast majority of India's urban population. The $8km$ MODIS markets also capture some of India's population that do not reside in Census' towns, which is why the share is above 100%.

The left axis of Figure 7 examines the correlation between population and area for $1km$ buffered markets. For each market definition, there is a strong positive correlation between the area of the market and its total population. This validates that the larger markets we detect are not simply capturing pixels that appear builtup but contain no population. Instead, larger markets contain more people, as we would expect. A second message of the graph is that the population variance across smaller landcover-based markets can be large. This again reflects the fact that distinct landcover based-markets can be found in both remote areas and large metropolises. (This variance decreases, but the positive correlation remains, at higher buffered $4km$ markets, as illustrated in Figure A5). We also examine population density, defined as population divided by land area, as a measure of economic activity on the right axis. The figure reveals a fairly constant density across

[14] An early version of this paper used WorldPop, a publicly available source of gridded population data. These data contain measurement errors but are nevertheless useful because of their global coverage. These figures are available in earlier versions of the paper and are available upon request.

size for each market definition. However, the figure does show higher variation in population density for smaller landcover markets for reasons just explained.

While the builtup pixels from daytime imagery undoubtedly contain man-made structures that do not necessarily contain human settlements (e.g., roads, freeway overpasses, dams, and power grids), the Census data serve as an important validation that the markets we identify do indeed contain urban populations within their boundaries.

3.2.2 Nightlight Intensity

Previous work by Henderson et al. (2012) and Henderson et al. (2018) demonstrate that nightlight intensity is a good proxy for national or regional GDP. Inspired by this work, we compare the average DN (nightlight intensity per unit of land) across markets. While the average DN for nightlight-based markets would be affected by blooming because of its impact on the extent of market boundaries, blooming is less of an issue for landcover-based markets since those boundaries are more accurately delineated.

Figure 8 reports the relationship between the average DN and the land area of a market ($1km$ buffers). For each of the landcover-based markets, larger markets are associated with higher DNs. Moreover, the change in DNs across market size is quite sharp. For example, a MODIS market at the $10th$ percentile of the area distribution has a mean nightlight intensity of 9.4 compared to a value of 27.2 at the $90th$ percentile. Henderson et al. (2012) report an elasticity of 0.3 for GDP with respect to DN, which implies that there is a GDP difference of 56.8% between markets that span the interdecile range of land area.

The figure also reveals that landcover-based markets exhibit more variance in DN intensity at smaller market sizes. For instance, for the smallest MODIS markets, we observe the full range of mean DNs (as seen by examining the range of points spanned along the y-axis for given points just to the right of the origin along the x-axis). This regularity is again a result of the fact that we detect small-in-area landcover-based markets both in remote regions of the country, where economic intensity is low (as indicated by low DNs), and within large urban centers, where DNs are high. This suggests that when using DN intensity as a proxy for the economic activity of landcover-based markets, researchers may want to account for the characteristics of the surrounding markets.

The correlations in Figure 8 thus suggest that the *pooling* of daytime and nightlight imagery may be a powerful means of characterizing the *combined* extensive and intensive margins of urban markets. While researchers interested in the economic geography of specific cities may want to bring information from external datasets, these correlations are promising for researchers interested in studying urban market activity at national or global scales.

3.2.3 Proximity to Infrastructure

A third way to examine whether our markets capture economic activity is to merge them with open-source data containing the locations of key infrastructure markers. We examine proximity of markets to paved roads, railway stations, and mobile phone towers. A caveat with this exercise is

that these data reflect the current location of infrastructure. The road and railway station data are from OpenStreetMaps.[15] Because the road data are for a time period roughly five years after our satellite imagery was collected, there is measurement error in matching markets to roads. Rail stations are less susceptible to this problem since they are built at much lower frequencies. The tower locations share the same caveat as the roads data, but have the advantage of being compiled by a different data source (https://opencellid.org).

We construct the distance between market centroids to the nearest infrastructure type for each market definition in Table 2. For nightlight-based $1km$ markets, the fractions of DN17.4, DN33 and DN60 markets that lie within two kilometers of a paved road are 96.7%, 97.0% and 97.4%, respectively. This fact should not be surprising since these markets are relatively large, although one caveat is that the nightlight data may capture street lights along the roads.

The more informative statistics are the fractions of landcover-based $1km$ markets that lie within two kilometers of a paved road. For MODIS market, this fraction is 88.3% (the corresponding numbers for GHSL and MIX markets are 89.4% and 90.8%). Since we believe that most urban markets would be connected to a road of some kind, this regularity provides validation that the daytime satellite imagery are capturing markets that contain economic activity. The table also reports proximity to the nearest railway station (second panel) and mobile towers (third panel).

We find that 26.2% of MODIS $1km$ markets are within $5km$ of a railway station, which rises to 81.7% for markets within $25km$ of a rail station. Proximity to mobile phone towers is also very high across markets: 86.9% of MODIS $1km$ markets are within five kilometers of a mobile tower.

We also expect a positive relationship between market size and its proximity to paved roads (Storeygard 2016). The left axis of Figure 9 plots this relationship, which illustrates the potential power of daytime imagery over nighttime imagery. Landcover-based markets exhibit a sharp negative elasticity of market area with respect to distance to the nearest road. For instance, compared to markets that are bisected by a road, a MODIS market that is $2km$ away from a road is about 50% smaller in land area. Such a large difference in size is not detectable using nightlight-based markets: for markets based on DN thresholds, the elasticity of size with respect to distance to a road is an imprecisely estimated zero.

Figure 9 repeats the plots with average nightlight intensity on the second y-axis. These illustrate that for landcover-based markets, light intensity, which as discussed above is a proxy for the intensity of economic activity, falls sharply with distance to a paved road. For MODIS markets, the average light value falls from about 20 to 8 when one compares a market that lies on top of a road to a market that is $2km$ from a road. As with land area, a decline in light intensity is not detectable for nightlight-based markets between $0km$-$2km$ from a road.

As noted earlier, nightlight data have a relatively coarse spatial resolution compared to daytime images ($1km$ vs $30m$). The lights data are also subject to blooming which introduces mea-

[15] We use the OpenStreetMaps road classifications. The major roads (511x) include motorways, freeways, and trunk, primary, secondary and tertiary roads. We additionally include two minor road classifications: smaller local roads (5121) and roads in residential areas (5112). For the railway stations, we include large rail stations (5601) and smaller, local rail stations or subway stations (5602).

surement error in market size. Which of these two differences—spatial resolution or exposure to blooming—explains why the road-distance elasticities are less sharply negative for nightlight-based markets when compared to landcover-based markets? We examine this question in the MODIS data by changing the minimum cluster threshold from 1 pixel to 4 pixels, or roughly $1km$ grid cells, in order to match the minimum market area of nightlight-based markets. We then rebuild the landcover-based markets using a $1km$ buffer. The procedure creates 5,527 markets (compared to 12,953 using a minimum of one MODIS pixels at $1km$ buffer). We then compare the elasticity of market area and average DN value to distance from the closest road in Appendix Figure A6. The MODIS markets that impose a $1km$ minimum area still display a strong negative elasticity with respect to road distance for both outcomes. With landcover-based markets and nightlight-based markets now approximately equal in spatial resolution, the more negative road-distance elasticity for the former relative to the latter would appear to be the result of blooming in nightlights and the measurement error it introduces when trying to detect market size.

4 Markets within Super-Markets

The literature has long recognized that actual structure of cities does not easily map into static spatial models with a featureless geography. Instead, urban sprawl occurs unevenly at city boundaries (Duranton and Puga 2014). As cities expand, there often remains undeveloped land within city limits. This may be due to physical constraints imposed by geography (Harari 2017), leapfrogging that occurs from dynamic city growth (Fujita 1982), municipalities wanting to control how land is utilized, or, particularly relevant to India, disputes over land titles and coordination failures across government agencies (Roy 2009). These features have led to a large literature on the polycentric structure of cities (Duranton and Puga 2015). We next explore this polycentricity.

4.1 Properties of Super-Markets

Our market definitions have a recursive property that nests smaller buffered markets within larger buffered "super-markets". This feature allows us to study the distribution of markets within super-markets. Our results suggest a potential use of high-resolution daytime satellite imagery to evaluate policies that impact the intra-regional distribution of markets within these larger urban forms. The granularity allows us to observe impacts both within markets (e.g., markets or neighborhoods within a larger super-market), and at high temporal frequencies (important for policymakers loathe to wait years to evaluate the returns to public infrastructure investments).

To see that landcover-based markets have the potential to uncover local-level responses to shocks that would otherwise appear hidden by the coarseness and granularity of nightlight-based markets, consider Figure 10, which maps MODIS landcover-based markets at different buffers for New Delhi. The outer ring is the official administrative boundary of New Delhi. The light gray polygon represents the 60 $8km$ buffered markets that lie within the administrative boundary. These $8km$ super-markets further contain smaller $4km$, $2km$ and $1km$ markets. Within the official

boundary, we detect 205, 435 and 579 markets buffered at $4km$, $2km$, and $1km$.

Turning to the country as a whole, Table 3 reports the average number of $i = \{1, 2, 4\}km$ markets that are contained within their larger super-market buffer $j = \{2, 4, 8\}km$ for all markets in India. While the megacity of New Delhi unsurprisingly stands out for its large number of markets, the presence of these markets is a general phenomenon detectable via landcover-based market definitions. For example, an average of 1.9 buffered MODIS $1km$ markets lie within super-markets defined at a $4km$ buffer, and an average of 4.2 markets lie within $8km$ super-markets. The second column within each panel of Table 3 reports the elasticity of the number of markets to the size of the super-market. The elasticity of the number of $1km$ MODIS markets with respect to the size of $2km$ markets is 0.15 and increases to 0.31 and 0.36 for $4km$ and $8km$ super-markets, respectively. These patterns suggests that there is substantial scope for using landcover-based markets to evaluate theories of how polycentric cities are organized and grow. Markets defined according to administrative boundaries would likely be poorly suited for this purpose as official boundary definitions may substantially lag urban structure.

The size of markets within super-markets is highly unequal. Table 4 reports the distribution of $1km$ market size shares within the super-markets for MODIS. For each $1km$ market, we rank them within their respective super-market and compute their share of builtup area. The top panel reports the distribution of shares within $2km$ super-markets; the middle and bottom panels reports statistics for $4km$ and $8km$ super-markets, respectively. The table reveals that for $4km$ super-markets that contain two $1km$ markets, the larger market accounts for about 72% of the builtup area. For super-markets that contain 5 markets, the largest market accounts for 51% of the builtup area of $4km$ super-markets, and 58% of the builtup area of $8km$ super-markets.

While super-markets contain many distinct markets, they also contain vast tracts of unbuilt land. To demonstrate this regularity, we compute the area of builtup pixels that lie within the boundary. Figure 5 plots a non-parametric relationship for the developed land fraction against the size of markets, by buffer. For the $1km$ and $2km$ buffers, a large fraction of market area is builtup for both nightlight and MODIS markets. This is intuitive since the clustering algorithm builds very small land bridges for these buffers. For larger buffers, the fraction of land area increases with area for nightlight markets. However, for MODIS markets, the builtup area percentage falls with area size; it levels off at around 50% for $4km$ markets, and falls continuously in $8km$ markets. The fraction of builtup land area in the average $8km$ buffered DN33 and DN60 market is 78% and 87%, respectively, compared to just 23% in the average $8km$ MODIS market.

These patterns reinforce several messages from earlier figures. The blooming of lights implies that larger buffered nightlight-based markets will suggest that human activity is too expansive within its boundaries. This is particularly an issue for the largest cities. Landcover-based markets instead reveal far more undeveloped land within boundaries. For large cities, daytime imagery reveal a sizable fraction of undeveloped land within market boundaries. This suggests there may be substantial within-market variation in builtup land across regions and over time.

4.2 Application to Market Access

How might we deploy data on landcover-based markets and super-markets? One application is to detect the consequences of infrastructure development. Governments across the developing world are making large-scale investments in improving internal transport connectivity. A growing literature studies the economic impacts of transportation (e.g, see Redding and Turner 2015). Ghani et al. (2014), for instance, use across-district variation in the distance to the India's Golden Quadrilateral highways and find positive impacts on allocative efficiency within Indian manufacturing. Asher and Novosad (forthcoming) study India's $40 billion in expenditures on rural roads and do not find substantial effects on rural household welfare. Both analyses draw upon administrative datasets to evaluate impacts of new roads. Satellite imagery offers the potential to complement these studies by using remotely sensed data and by analyzing impacts on markets that lie within, for instance, larger buffered peri-urban areas.

In the spirit of such analysis, we examine the average distances to (the centroids of) other markets within given super-markets, which are reported in third column of each panel in Table 3. Consider MODIS markets. Within a $8km$ buffer, the average distance between $1km$ sub-markets is 52.1 kilometers, indicating that the typical $8km$ buffered super-market is an economic region unto itself, which would utilize highways and railways in a manner that we may typically associate with inter-urban transport. The average distance between $1km$ sub-markets within a $4km$ buffer is 6.2 kilometers, which indicates that at a $4km$ buffer we are dealing with collections of interconnected neighborhoods. The contrast in market distances between $4km$ and $8km$ buffered super-markets illustrates the different market concepts that these designations represent. One might reasonably conclude that $4km$ buffered markets approximately constitute commuting zones, while $8km$ buffered markets approximately constitute economic regions that support dense internal trade in goods and services. Differing urban market definitions may then be useful for evaluating the consequences of reduced travel time on different aspects of economic integration, for goods markets at higher distance buffers and for local labor markets at lower distance buffers.

To investigate such potential, we follow Donaldson and Hornbeck (2016) by calculating measures of market access for the MODIS markets. For each market i, we calculate its market access as

$$MA_i = \sum_{j \in S_{ik}, j \neq i} \frac{area_j}{distance_{ij}^{\theta}} + \sum_{j \notin S_{ik}, j \neq i} \frac{area_j}{distance_{ij}^{\theta}} \tag{1}$$

where $area_j$ is the land area of market j, $distance_{ij}$ is the great circle distance from market i to market j, and θ is a distance elasticity that we set to 1.4 (Redding and Turner, 2015). We exclude the own market in the summation, as Donaldson and Hornbeck (2016) do in their analysis. We are particularly interested in the contribution to market i's term by the j markets that lie within i's super-market S_{ik}, buffered at $k = \{2, 4, 8\}km$. Figure 11 reports the contribution of the within-super-market component, across buffers and daytime imagery sources. We also report the results that obtain for other distance elasticities by setting $\theta = 1$ and $\theta = 1.8$.

In the baseline case of $\theta = 1.4$, the results indicate that, on average, 2.3%, 6.8% and 25.4% of

a $1km$ MODIS market's access comes from other markets within the same super-market buffer of $2km$, $4km$, and $8km$, respectively. At the higher elasticity of $\theta = 1.8$, the corresponding percentages increase to 5.2%, 14.4% and 40.1%. Whereas previous literature largely conceives of infrastructure development as integrating our equivalent of super-markets, examining landcover-based markets reveals that a substantial share of a location's market access is intra-urban in nature. With data on combined infrastructure investments in inter-state highways, such as India's Golden Quadrilateral, and in intra-urban investments in access roads, road widening, and related improvements, daytime satellite imagery have the potential to provide a much higher resolution characterization of how these changes in trade costs shape the spatial distribution of economic activity.

5 Conclusion

Economists have been utilizing satellite imagery for over a decade. Notable applications have elucidated the dimensions of urban sprawl and the connection between GDP growth and the intensity of light emitted at night. In the last several years, the landscape, so to speak, has begun to change rapidly. Dramatic reductions in storage costs have made vast troves of high-resolution daytime satellite imagery widely available, while advances in machine learning are making it possible to deploy imagery to detect economic outcomes at previously unimaginable spatial resolutions.

Our results indicate the value of combining different types of satellite imagery in economic analysis. Daytime imagery is well suited for defining the spatial expanse of markets, the polycentricity of urban areas, and the gaps in urban development that exist even within densely populated cities. Nighttime imagery, in turn, is well suited for measuring the intensive margin of economic activity within urban ares.

The creation of new methods for integrating alternative sources of satellite imagery is a promising avenue for research.With existing analytical tools, these data will make it possible to evaluate the potentially highly spatially heterogeneous economic impacts of investments in infrastructure and other policy interventions. With the continents of Asia and Africa in the midst of a multi-trillion dollar infrastructure investments, the arrival of such capabilities is well timed.

Although satellite imagery greatly expands the supply of data amenable to economic analysis, their interpretation is, at this stage, still constrained by the supply of conventionally measured economic quantities, which serve as ground truth in machine learning. Demand will be particularly high for methods to validate satellite-based measures of economic activity using additional sources of micro data. We view this as an important area for future work.

References

Abrahams, A., C. Oram, and N. Lozano-Gracia (2018): "Deblurring DMSP nighttime lights: A new method using Gaussian filters and frequencies of illumination," *Remote Sensing of Environment*, 210, 242 – 258.

Alqurashi, A. F. and L. Kumar (2013): "Investigating the use of remote sensing and GIS techniques to detect land use and land cover change: A review," *Advances in Remote Sensing*, 2, 193.

Anas, A., R. Arnott, and K. A. Small (1998): "Urban spatial structure," *Journal of Economic Literature*, 36, 1426–1464.

Angel, S., J. Parent, and D. L. Civco (2010): "Ten compactness properties of circles: measuring shape in geography," *The Canadian Geographer/Le Géographe canadien*, 54, 441–461.

Asher, S. and P. Novosad (forthcoming): "Rural Roads and Local Economic Development," *American Economic Review*.

Bagan, H. and Y. Yamagata (2015): "Analysis of urban growth and estimating population density using satellite images of nighttime lights and land-use and population data," *GIScience & Remote Sensing*, 52, 765–780.

Bleakley, H. and J. Lin (2012): "Portage and path dependence," *The quarterly journal of economics*, 127, 587–644.

Burchfield, M., H. Overman, D. Puga, and M. Turner (2006): "Causes of Sprawl: A Portrait from Space," *The Quarterly Journal of Economics*, 121, 587–633.

Cauwels, P., N. Pestalozzi, and D. Sornette (2014): "Dynamics and spatial distribution of global nighttime lights," *EPJ Data Science*, 3, 2.

Channan, S., K. Collins, and W. R. Emanuel (2014): "Global mosaics of the standard MODIS Vegetation Continuous Fields data," University of Maryland and the Pacific Northwest National Laboratory, College Park, Maryland, USA.

Davis, D. R., J. I. Dingel, and A. Miscio (2018): "Cities, Skills, and Sectors in Developing Economies," *mimeo Columbia University*.

de Bellefon, M.-P., P.-P. Combes, G. Duranton, and L. Gobillon (2018): "Delineating Urban Areas Using Building Density," mimeo the wharton school, mimeo The Wharton School.

Desmet, K. and J. V. Henderson (2015): "Chapter 22 - The Geography of Development Within Countries," in *Handbook of Regional and Urban Economics*, ed. by G. Duranton, J. V. Henderson, and W. C. Strange, Elsevier, vol. 5 of *Handbook of Regional and Urban Economics*, 1457 – 1517.

Donaldson, D. and R. Hornbeck (2016): "Railroads and American Economic Growth: A Market Access Approach," *The Quarterly Journal of Economics*, 131, 799–858.

Donaldson, D. and A. Storeygard (2016): "The View from Above: Applications of Satellite Data in Economics," *Journal of Economic Perspectives*, 30, 171–98.

Duranton, G. (2015): "A Proposal to Delineate Metropolitan Areas in Colombia," 2015, 223–264.

Duranton, G. and D. Puga (2001): "Nursery Cities: Urban Diversity, Process Innovation, and the Life Cycle of Products," *American Economic Review*, 91, 1454–1477.

——— (2004): "Chapter 48 - Micro-Foundations of Urban Agglomeration Economies," in *Cities and Geography*, ed. by J. V. Henderson and J.-F. Thisse, Elsevier, vol. 4 of *Handbook of Regional and Urban Economics*, 2063 – 2117.

——— (2014): "Chapter 5 - The Growth of Cities," in *Handbook of Economic Growth*, ed. by P. Aghion and S. N. Durlauf, Elsevier, vol. 2 of *Handbook of Economic Growth*, 781 – 853.

——— (2015): "Chapter 8 - Urban Land Use," in *Handbook of Regional and Urban Economics*, ed. by G. Duranton, J. V. Henderson, and W. C. Strange, Elsevier, vol. 5 of *Handbook of Regional and Urban Economics*, 467 – 560.

Duranton, G. and M. A. Turner (2012): "Urban Growth and Transportation," *The Review of Economic Studies*, 79, 1407–1440.

Eeckhout, J. (2004): "Gibrat's Law for (All) Cities," *American Economic Review*, 94, 1429–1451.

Elvidge, C. D., K. Baugh, M. Zhizhin, F. C. Hsu, and T. Ghosh (2017): "VIIRS night-time lights," *International Journal of Remote Sensing*, 38, 5860–5879.

Elvidge, C. D., K. E. Baugh, J. B. Dietz, T. Bland, P. C. Sutton, and H. W. Kroehl (1999): "Radiance Calibration of DMSP-OLS Low-Light Imaging Data of Human Settlements," *Remote Sensing of Environment*, 68, 77 – 88.

Friedl, M. and D. Sulla-Menashe (2015): "MCD12Q1 MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid V006," distributed by nasa eosdis land processes daac, distributed by NASA EOSDIS Land Processes DAAC.

Fu, H., Z. Shao, P. Fu, and Q. Cheng (2017): "The Dynamic Analysis between Urban Night-time Economy and Urbanization Using the DMSP/OLS Nighttime Light Data in China from 1992 to 2012," *Remote Sensing*, 9, 416.

Fujita, M. (1982): "Spatial patterns of residential development," *Journal of Urban Economics*, 12, 22 – 52.

Fujita, M., P. R. Krugman, and A. J. Venables (2001): *The spatial economy: Cities, regions, and international trade*, MIT press.

Gabaix, X. and R. Ibragimov (2011): "Rank - 1/2: A Simple Way to Improve the OLS Estimation of Tail Exponents," *Journal of Business and Economic Statistics*, 29, 24–39.

Gabaix, X. and Y. M. Ioannides (2004): "The evolution of city size distributions," in *Handbook of regional and urban economics*, Elsevier, vol. 4, 2341–2378.

Gennaioli, N., R. L. Porta, F. L. de Silanes, and A. Shleifer (2013): "Human Capital and Regional Development," *The Quarterly Journal of Economics*, 128, 105–164.

Ghani, E., A. G. Goswami, and W. R. Kerr (2014): "Highway to Success: The Impact of the Golden Quadrilateral Project for the Location and Performance of Indian Manufacturing," *The Economic Journal*, 126, 317–357.

Goldblatt, R., M. Stuhlmacher, B. Tellman, N. Clinton, G. Hanson, M. Georgescu, C. Wang, F. Serrano-Candela, A. Khandelwal, W. Cheng, and R. Balling (2018): "Using Landsat and nighttime lights for supervised pixel-based image classification of urban land cover," *Remote Sensing of Environment*, 205, 253–275.

Goldblatt, R., W. You, G. Hanson, and A. Khandelwal (2016): "Detecting the Boundaries of Urban Areas in India: A Dataset for Pixel-Based Image Classification in Google Earth Engine," *Remote Sensing*, 8, 634.

Guo, W., D. Lu, Y. Wu, and J. Zhang (2015): "Mapping impervious surface distribution with integration of SNNP VIIRS-DNB and MODIS NDVI data," Remote Sensing.

Hanson, G. (2005): "Market potential, increasing returns and geographic concentration," *Journal of International Economics*, 67, 1–24.

Harari, M. (2017): "Cities in bad shape: Urban geometry in india," .

Henderson, J. V. (1974): "The Sizes and Types of Cities," *American Economic Review*, 64, 640–56.

Henderson, J. V., T. Squires, A. Storeygard, and D. Weil (2018): "The Global Distribution of Economic Activity: Nature, History, and the Role of Trade1," *The Quarterly Journal of Economics*, 133, 357–406.

Henderson, J. V., A. Storeygard, and D. N. Weil (2012): "Measuring Economic Growth from Outer Space," *American Economic Review*, 102, 994–1028.

Henderson, V. and A. Mitra (1996): "The new urban landscape: Developers and edge cities," *Regional Science and Urban Economics*, 26, 613–643.

Huang, Q., C. He, B. Gao, Y. Yang, Z. Liu, Y. Zhao, and Y. Dou (2015): "Detecting the 20 year city-size dynamics in China with a rank clock approach and DMSP/OLS nighttime data," *Landscape and Urban Planning*, 137, 138–148.

Huang, X., A. Schneider, and M. Friedl (2016): "Mapping sub-pixel urban expansion in China using MODIS and DMSP/OLS nighttime lights," Remote Sensing of Environment.

Jean, N., M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon (2016): "Combining satellite imagery and machine learning to predict poverty," *Science*, 353, 790–794.

Mertes, C., A. Schneider, D. Sulla-Menashe, A. Tatem, and B. Tan (2015): "Detecting change in urban areas at contintental scales with MODIS data," Remote Sensing of Environment.

Michaels, G., D. Nigmatulina, F. Rauch, T. Regan, N. Baruah, and A. Dahlstrand-Rudin (2018): "Planning ahead for better neighborhoods: long run evidence from Tanzania," Tech. rep., London School of Economics.

Michalopoulos, S. and E. Papaioannou (2013a): "National institutions and subnational development in Africa," *The Quarterly Journal of Economics*, 129, 151–213.

——— (2013b): "Pre-colonial ethnic institutions and contemporary African development," *Econometrica*, 81, 113–152.

Otsu, N. (1979): "A Threshold Selection Method from Gray-Level Histograms," *IEEE Transactions on Systems, Man and Cybernetics*, 9, 62–66.

Pandey, B., P. Joshi, and K. C. Seto (2013): "Monitoring urbanization dynamics in India using DMSP-OLS night time lights and SPOT-VGT data," *International Journal of Applied Earth Observation and Geoinformation*, 23, 49–61.

Pesaresi, M., D. Ehrilch, A. J. Florczyk, S. Freire, A. Julea, T. Kemper, P. Soille, and V. Syrris (2015): "GHS built-up confidence grid, derived from Landsat, multitemporal (1975, 1990, 2000, 2014)," European Commission, Joint Research Centre (JRC).

Pinkovskiy, M. and X. Sala-i Martin (2016): "Lights, Camera... Income: Illuminating the National Accounts- Household Surveys Debate," *The Quarterly Journal of Economics*, 131, 579–631.

Redding, S. J. (2016): "Goods trade, factor mobility and welfare," *Journal of International Economics*, 101, 148 – 167.

Redding, S. J. and M. A. Turner (2015): "Chapter 20 - Transportation Costs and the Spatial Organization of Economic Activity," in *Handbook of Regional and Urban Economics*, ed. by G. Duranton, J. V. Henderson, and W. C. Strange, Elsevier, vol. 5 of *Handbook of Regional and Urban Economics*, 1339 – 1398.

Roy, A. (2009): "Why India Cannot Plan Its Cities: Informality, Insurgence and the Idiom of Urbanization," *Planning Theory*, 8, 76–87.

Rozenfeld, H. D., D. Rybski, X. Gabaix, and H. A. Makse (2011): "The Area and Population of Cities: New Insights from a Different Perspective on Cities," *American Economic Review*, 101, 2205–25.

Shi, K., C. Huang, B. Yu, B. Yin, Y. Huang, and J. Wu (2014): "Evaluation of NPP-VIIRS night-time light composite data for extracting built-up urban areas," *Remote Sensing Letters*, 5, 358–366.

Song, X.-P., J. O. Sexton, C. Huang, S. Channan, and J. R. Townshend (2016): "Characterizing the magnitude, timing and duration of urban growth from time series of Landsat-based estimates of impervious cover," *Remote Sensing of Environment*, 175, 1 – 13.

Storeygard, A. (2016): "Farther on down the road: transport costs, trade and urban growth in sub-Saharan Africa," *The Review of economic studies*, 83, 1263–1295.

Tolbert, C. M. and M. Sizer (1996): "US Commuting Zones and Labor Market Areas: A 1990 Update," *U.S. Department of Agriculture*, Economic Research Service Staff Paper.

Trianni, G., G. Lisini, E. Angiuli, E. Moreno, P. Dondi, A. Gaggia, and P. Gamba (2015): "Scaling up to national/regional urban extent mapping using Landsat data," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 8, 3710–3719.

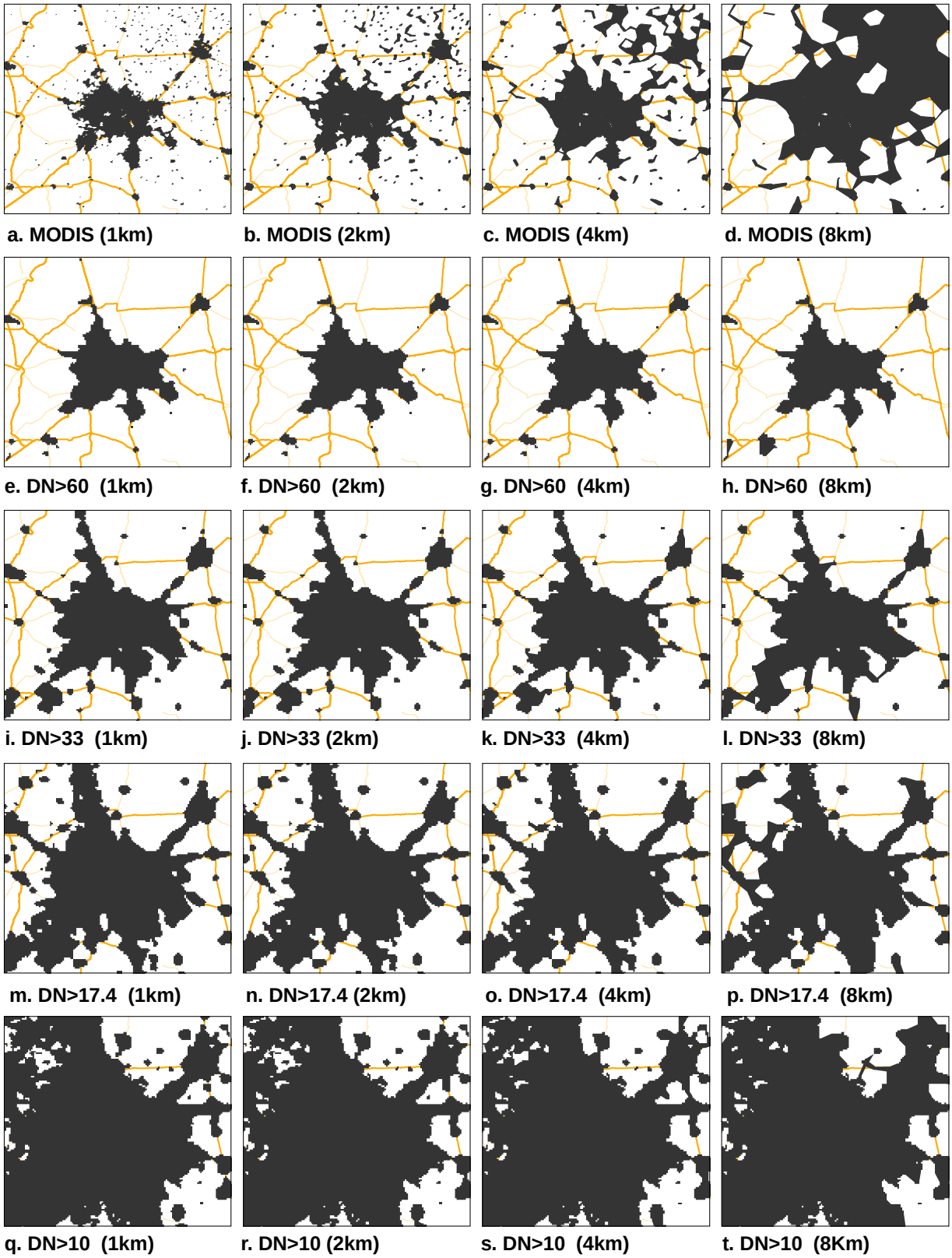
Tuttle, B. T., S. Anderson, C. Elvidge, T. Ghosh, K. Baugh, and P. Sutton (2014): "Aladdin's Magic Lamp: Active Target Calibration of the DMSP OLS," *Remote Sensing*, 6, 12708–12722.

Zhou, Y., S. J. Smith, C. D. Elvidge, K. Zhao, A. Thomson, and M. Imhoff (2014): "A cluster-based method to map urban area from DMSP/OLS nightlights," *Remote Sensing of Environment*, 147, 173–185.

Zhou, Y., S. J. Smith, K. Zhao, M. Imhoff, A. Thomson, B. Bond-Lamberty, G. R. Asrar, X. Zhang, C. He, and C. D. Elvidge (2015): "A global map of urban extent from nightlights," *Environmental Research Letters*, 10, 054011.

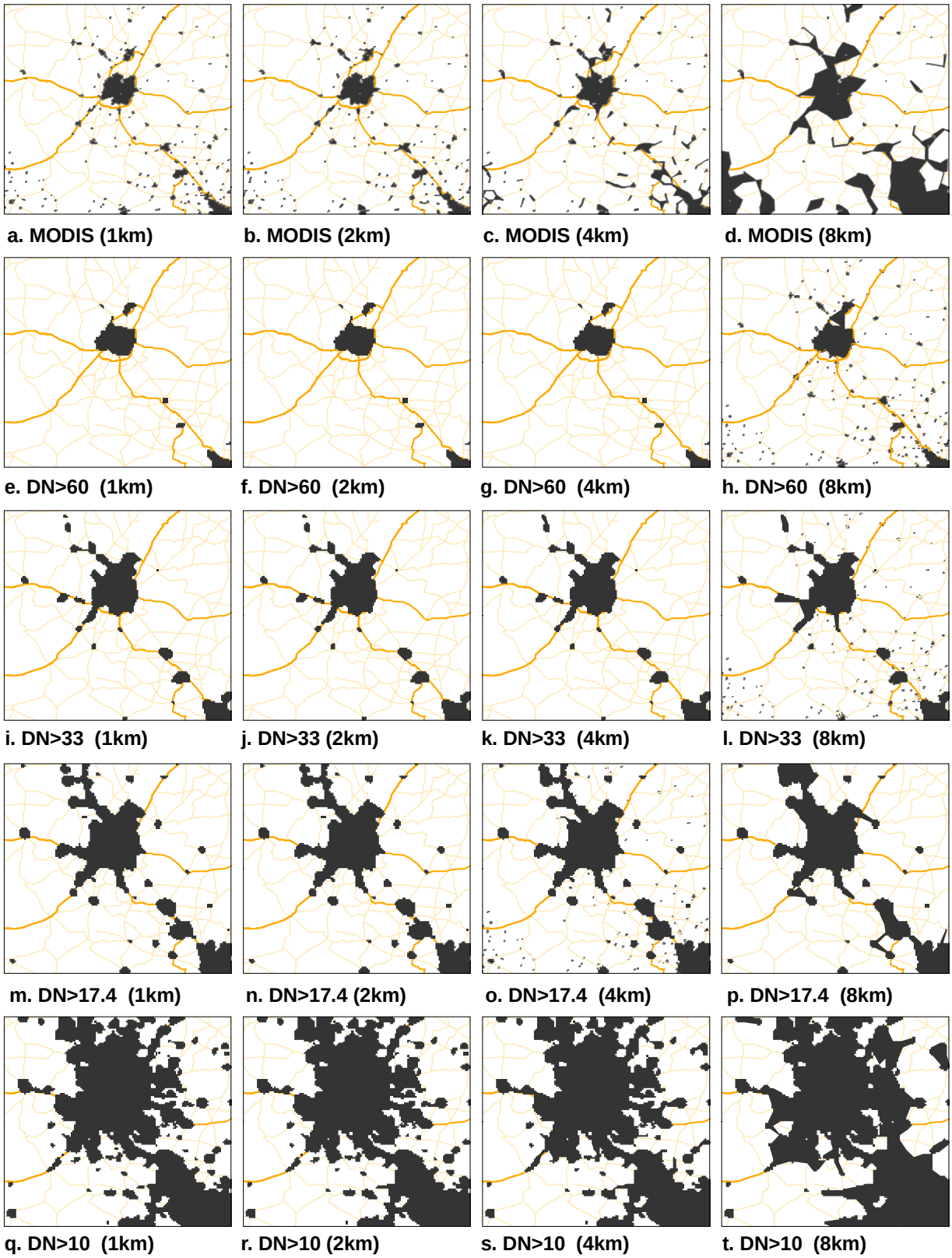
Figures and Tables

Figure 1: Delhi, Alternative Market Definitions



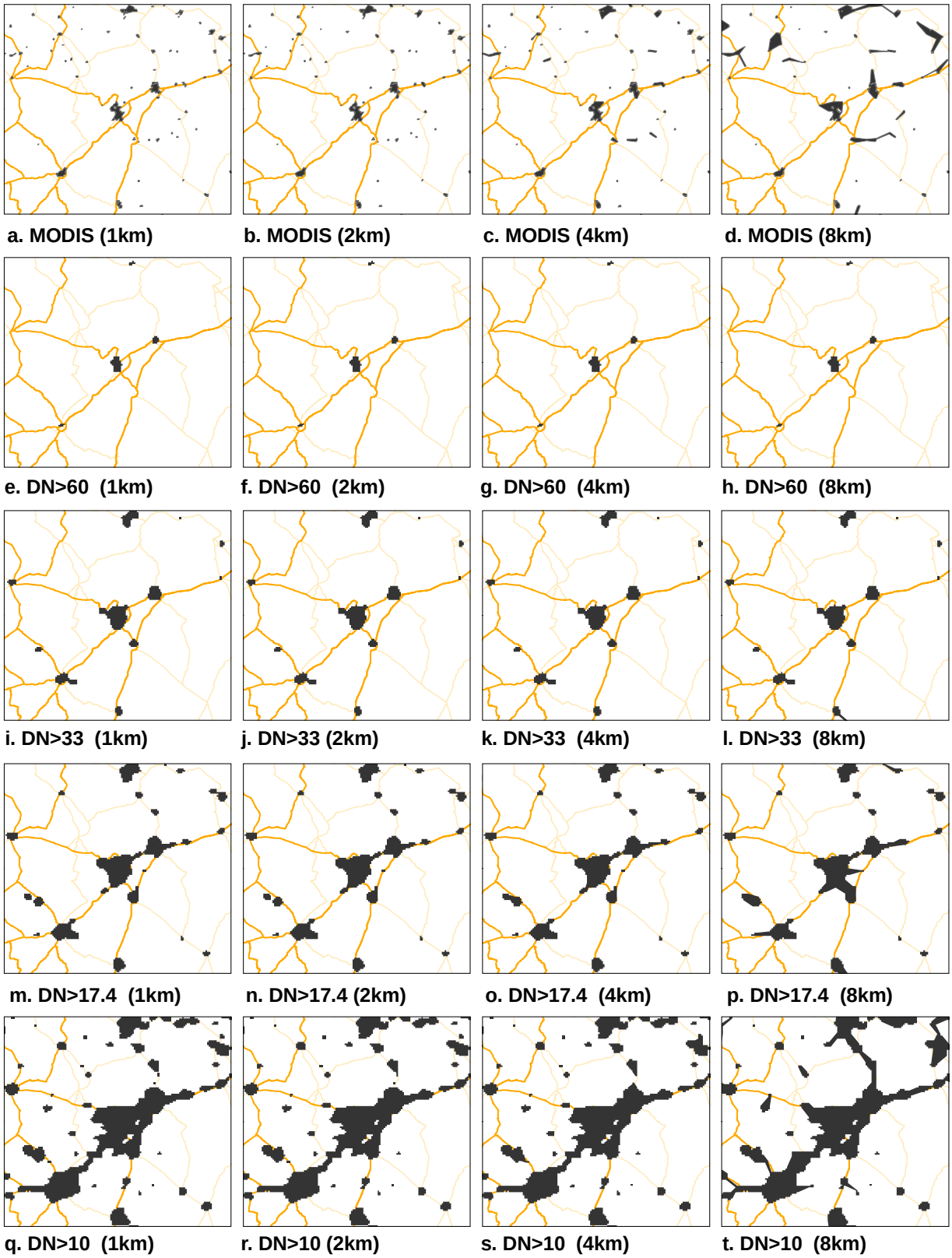
Notes: The figure displays markets around New Delhi for alternative distance buffers. Row 1 displays landcover-based markets using the MODIS layer. Row 2-5 displays nightlight-based markets.

Figure 2: Ahmedabad, Alternative Market Definitions



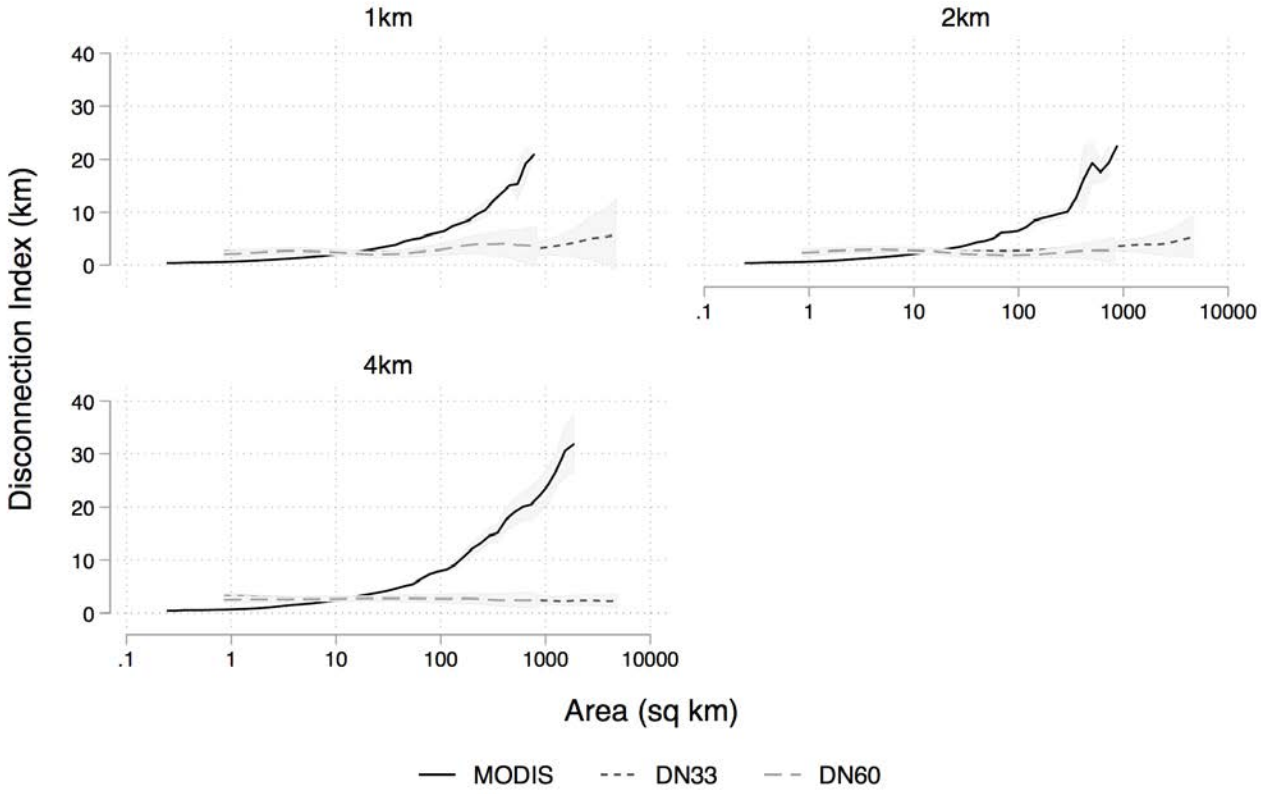
Notes: The figure displays markets around Ahmedabad for alternative distance buffers. Row 1 displays landcover-based markets using the MODIS layer. Row 2-5 displays nightlight-based markets.

Figure 3: Ajmer, Alternative Market Definitions



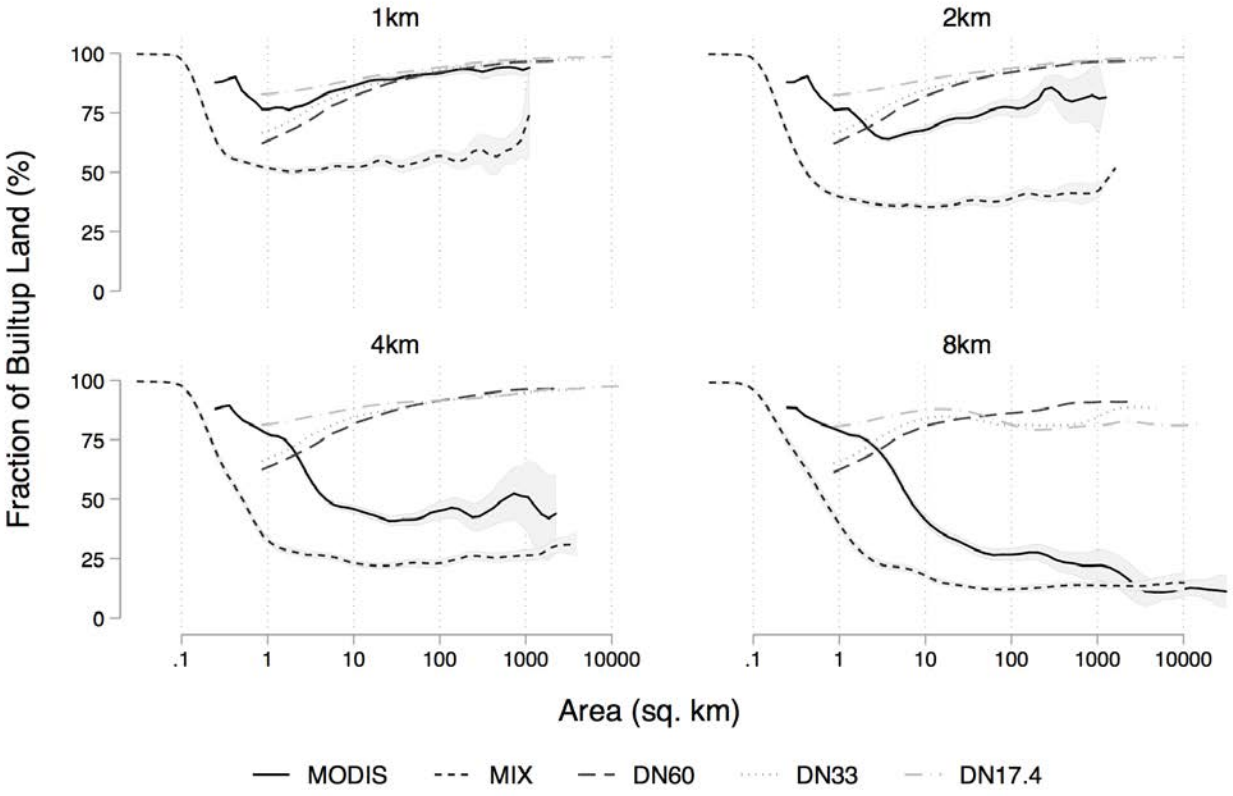
Notes: The figure displays markets around Ajmer for alternative distance buffers. Row 1 displays landcover-based markets using the MODIS layer. Row 2-5 displays nightlight-based markets.

Figure 4: Disconnection Index



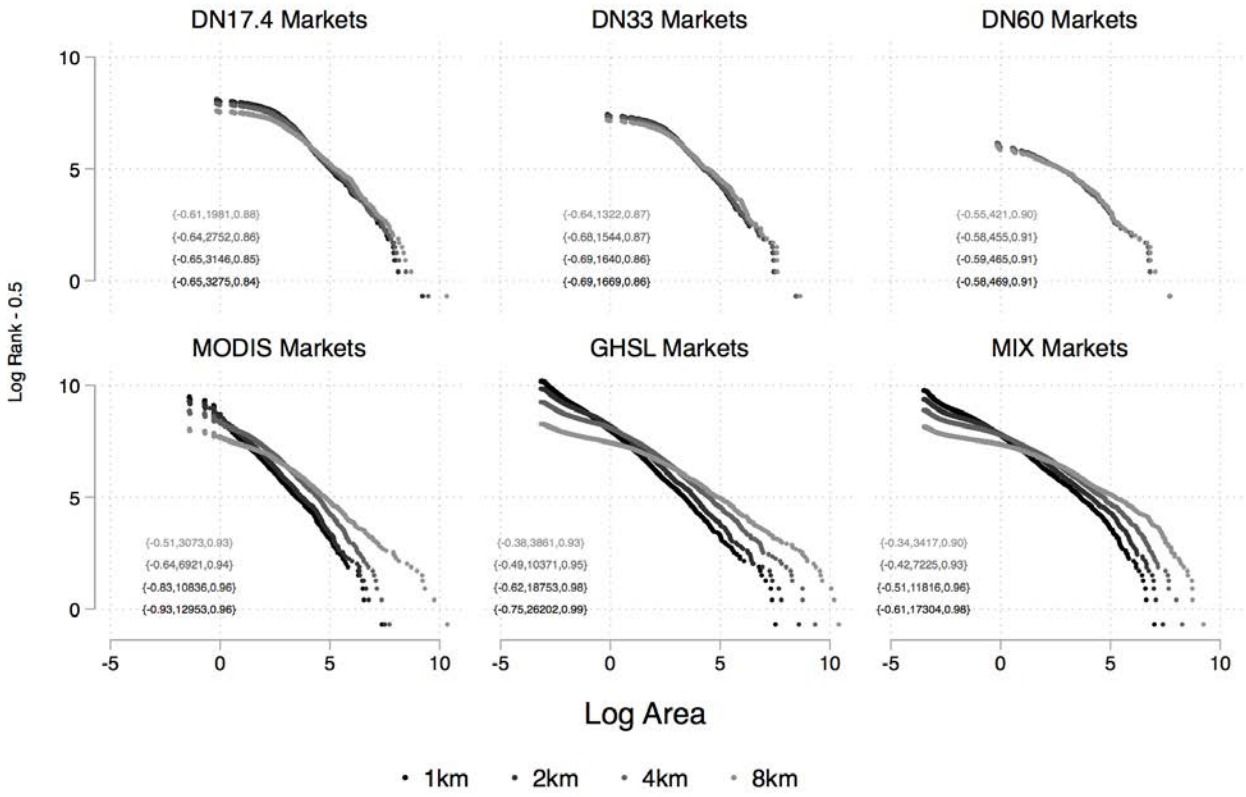
Notes: Figure reports a non-parametric plot between a market's disconnection index, measured in kilometers, and its area.

Figure 5: Builtup Land



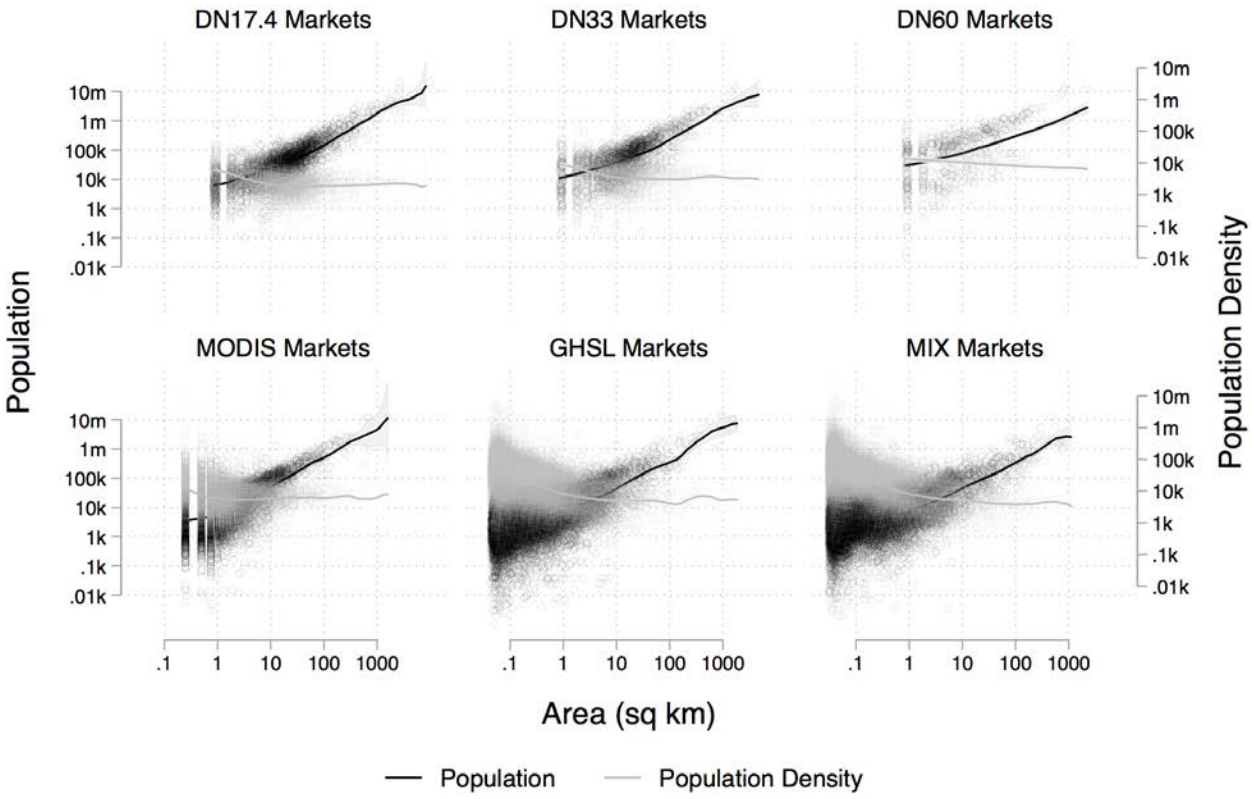
Notes: Figure reports the fraction of builtup land area by buffer. The nonparametric curve for MODIS markets displays 5%/95% confidence interval.

Figure 6: Land Area-Rank Relationship



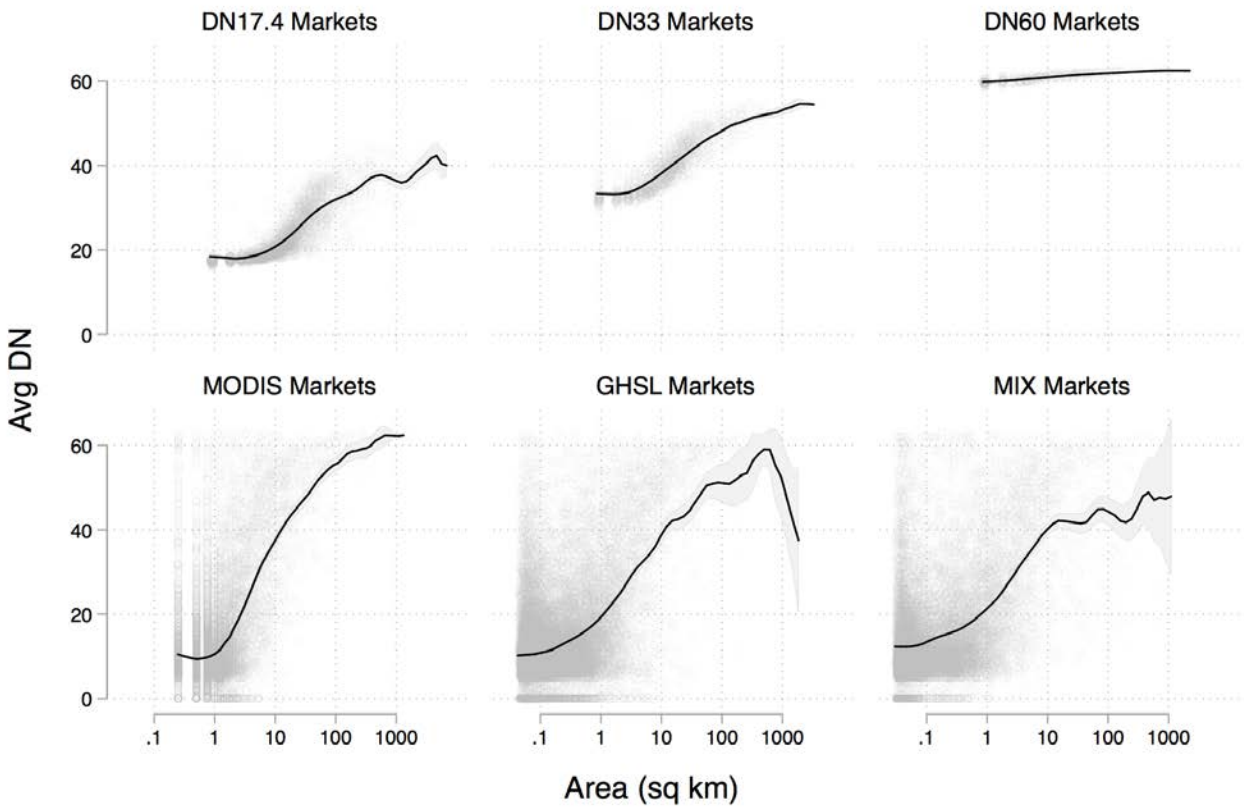
Notes: The {b, N, R-squared} are reported for the regression: log(rank-0.5) = constant + b*log(area) + error.

Figure 7: Population versus Land Area



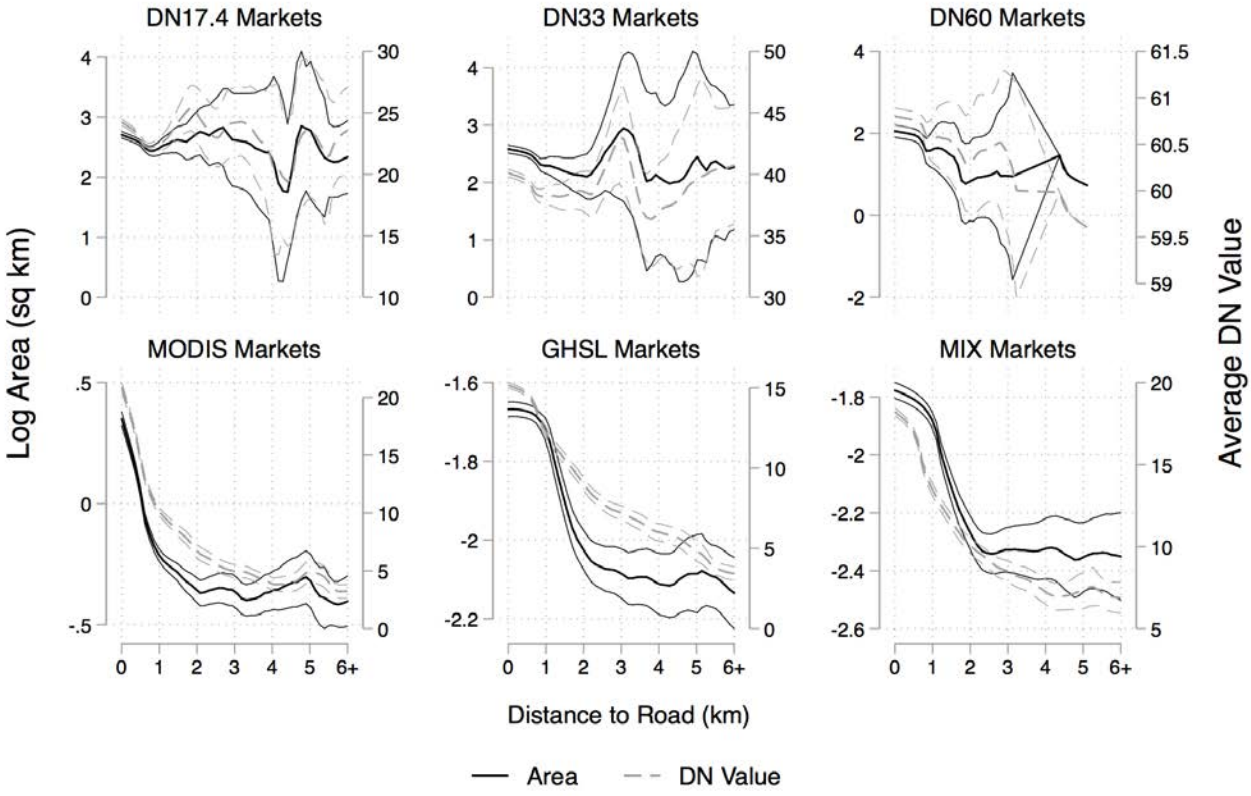
Notes: Figures report the relationship between market size, population and population density. Markets are buffered at $1km$. Population from 2011 Census.

Figure 8: Average DN Intensity versus Land Area



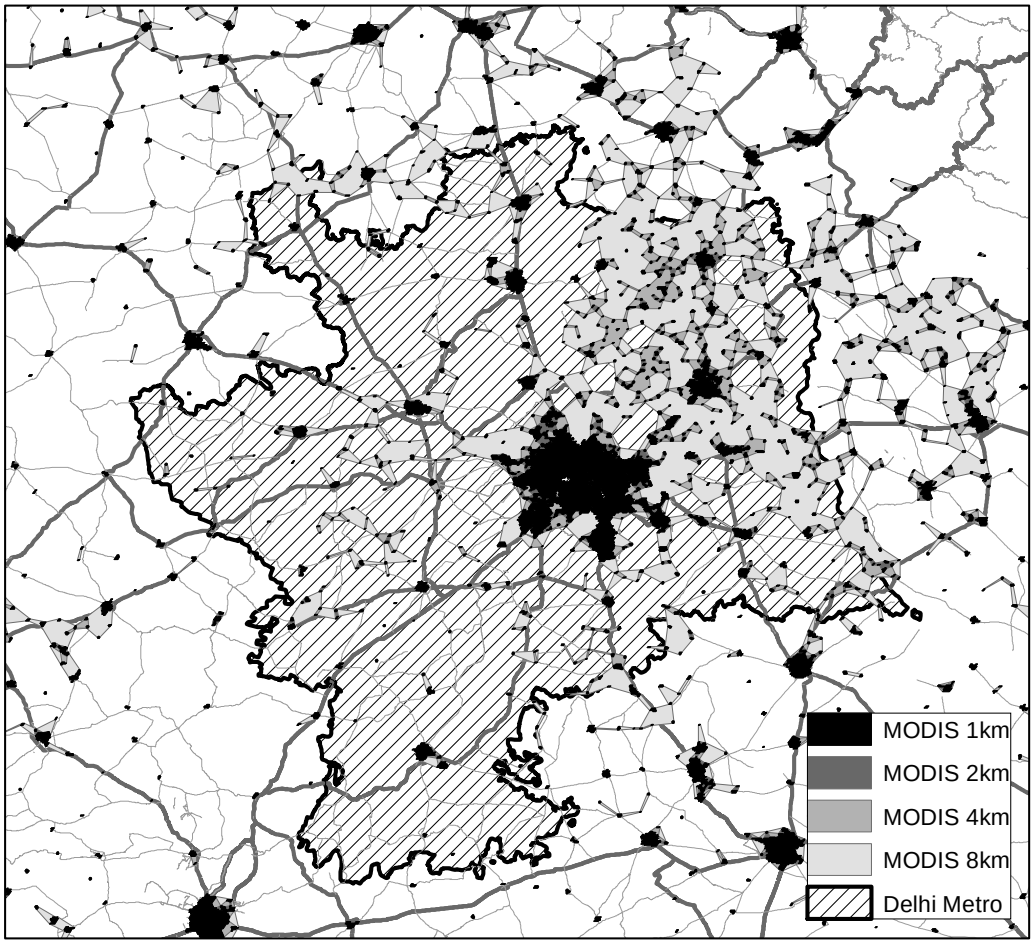
Notes: Figures report the relationship between market size and average light intensity. Markets are buffered at $1km$.

Figure 9: Land Area, Average DN and Proximity to Roads



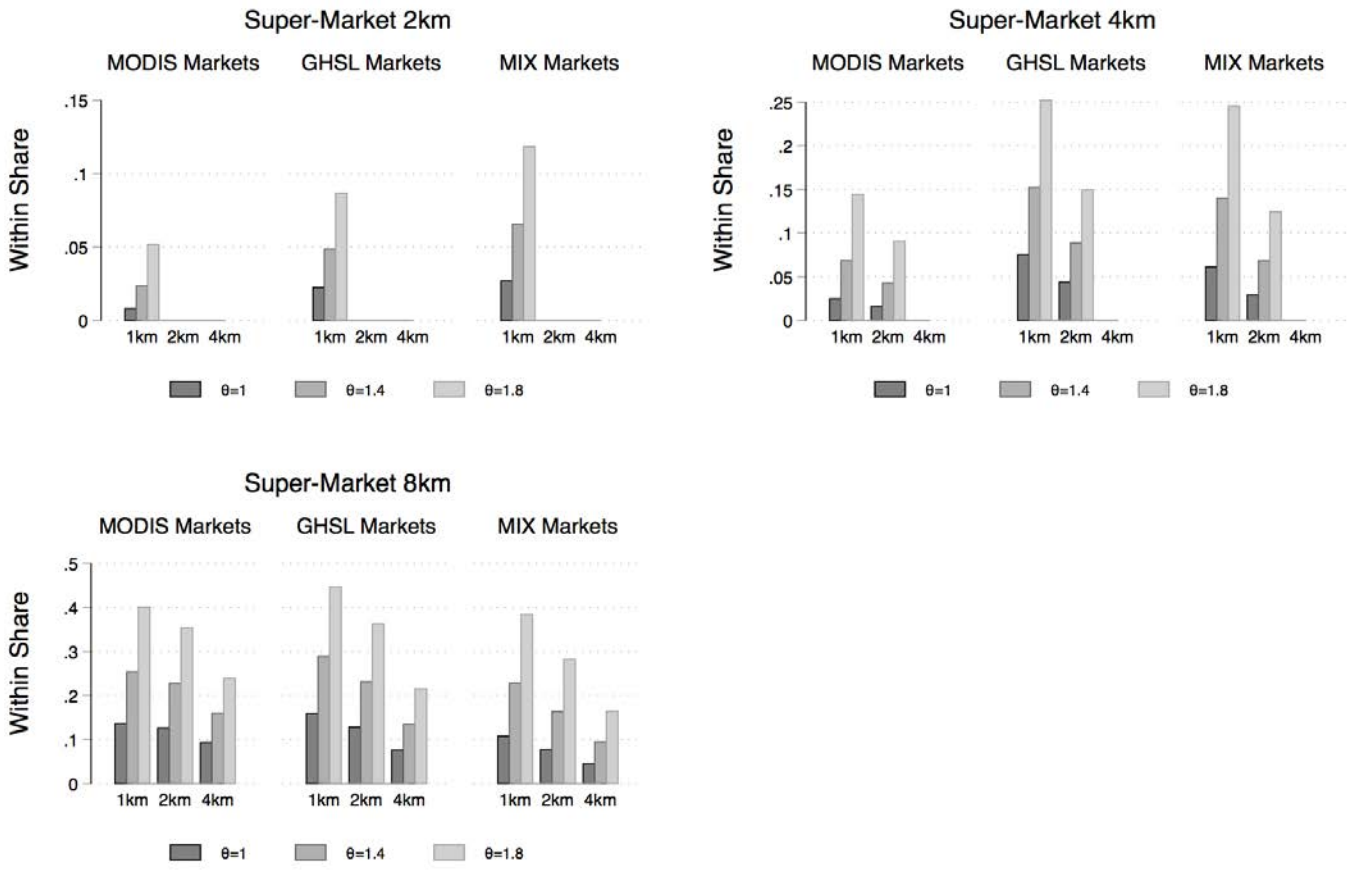
Notes: Distance to road is the shortest distance from market centroid to a primary, secondary or tertiary road. Road data obtained from OpenStreetMaps. Markets are buffered at $1km$. Figure shows 5% and 95% confidence intervals.

Figure 10: MODIS Landcover-Based Markets within New Delhi Metro Area



Notes: Map shows MODIS markets in the New Delhi metropolitan area. The black outline is the official administrative boundary of New Delhi from the 2011 Census. Within the administrative boundary, there are 579 $1km$, 435 $2km$, 205 $4km$ and 60 $8km$ markets.

Figure 11: Share of Market Access within Super-Markets



Notes: Figure reports the average share of market access accounted by markets within super-markets for different values of θ.

Table 1: Market Statistics

Market	Number	Avg Area (km^2)	Population Share	Urban Population Share
	Panel A: Nightlight-based Markets			
DN17.4				
1km	3,275	48.6	32.6%	104.6%
2km	3,275	50.7	32.6%	104.6%
4km	3,146	59.6	32.7%	105.0%
8km	2,752	97.8	33.5%	107.6%
DN33				
1km	1,669	39.0	23.4%	75.3%
2km	1,640	39.8	23.4%	75.3%
4km	1,544	42.9	23.5%	75.5%
8km	1,322	55.4	23.8%	76.5%
DN60				
1km	469	37.0	14.8%	47.6%
2km	465	37.3	14.8%	47.6%
4km	455	38.3	14.8%	47.6%
8km	421	43.7	14.9%	47.7%
	Panel B: Landcover-based Markets			
MODIS				
1km	12,953	3.0	29.0%	93.2%
2km	10,836	4.2	29.2%	93.8%
4km	6,921	10.6	30.1%	96.7%
8km	3,073	63.4	34.6%	111.1%
GHSL				
1km	26,202	1.4	33.3%	106.9%
2km	18,753	2.9	33.5%	107.6%
4km	10,371	10.9	34.8%	111.8%
8km	3,861	77.5	39.4%	126.5%
MIX				
1km	17,304	1.9	27.1%	87.1%
2km	11,816	4.3	27.3%	87.7%
4km	7,225	12.1	28.4%	91.1%
8km	3,417	54.5	31.4%	100.7%

Notes: Table reports the number and average area (in square kilometers) of markets and share of total India population, by definition. Total 2011 India population is 1,210,854,977. Urban population (population that resides in Census "Towns") is 377,106,125.

Table 2: Market Distances to Nearest Infrastructure

	Road					
Market	1km	2km	5km	10km	25km	50km
DN17.4	91.9%	96.7%	98.2%	98.6%	98.7%	98.7%
DN33	93.2%	97.0%	98.6%	99.0%	99.0%	99.0%
DN60	94.7%	97.4%	98.7%	98.9%	98.9%	98.9%
MODIS	75.1%	88.3%	97.3%	99.1%	99.3%	99.3%
GHSL	81.3%	89.4%	97.1%	99.0%	99.2%	99.2%
MIX	81.0%	90.8%	97.8%	99.1%	99.3%	99.3%
	Rail Station					
	1km	2km	5km	10km	25km	50km
DN17.4	12.2%	28.5%	43.1%	55.0%	83.4%	97.3%
DN33	19.2%	42.7%	60.8%	70.1%	89.2%	98.6%
DN60	22.0%	52.9%	78.7%	88.1%	97.0%	99.1%
MODIS	4.6%	12.8%	26.2%	46.2%	81.7%	96.8%
GHSL	5.1%	9.1%	22.5%	45.2%	82.1%	96.9%
MIX	6.1%	11.3%	26.7%	50.0%	83.6%	97.3%
	Mobile Phone Towers					
	1km	2km	5km	10km	25km	50km
DN17.4	59.7%	61.6%	64.4%	67.6%	69.9%	70.0%
DN33	96.6%	97.7%	99.0%	99.7%	100.0%	100.0%
DN60	98.9%	99.4%	99.8%	100.0%	100.0%	100.0%
MODIS	56.1%	68.2%	86.9%	96.8%	99.9%	100.0%
GHSL	55.3%	67.8%	86.8%	97.0%	99.9%	100.0%
MIX	59.6%	72.2%	89.6%	97.8%	99.9%	100.0%

Notes: Table reports the fraction of markets in which the centroid lies within a particular distance of the noted infrastucture type.

Table 3: Markets within Super-Markets

	2km Super-Market			4km Super-Market			8km Super-Market		
Market	Number	Elasticity	Distance	Number	Elasticity	Distance	Number	Elasticity	Distance
MODIS									
1km	1.2	0.15%	1.9	1.9	0.31%	6.2	4.2	0.36%	52.1
2km				1.6	0.23%	5.0	3.5	0.32%	50.4
4km							2.3	0.24%	38.1
GHSL									
1km	1.4	0.19%	5.1	2.5	0.28%	20.8	6.8	0.32%	75.4
2km				1.8	0.20%	13.5	4.9	0.28%	66.3
4km							2.7	0.22%	45.0
MIX									
1km	1.5	0.17%	3.9	2.4	0.25%	11.1	5.1	0.30%	31.6
2km				1.6	0.17%	7.2	3.5	0.25%	27.9
4km							2.1	0.18%	20.0

Notes: Table reports statistics for the $2km$, $4km$ and $8km$ super-markets. Columns 1, 4 and 7 are the average number of sub-markets within the super-market. Column 2, 5 and 8 is the average distance between sub-markets. Column 3, 6 and 9 is the elasticity of the number of sub-markets to the size of the super-market (e.g., a one percent increase in the size of the super-market increases the number of markets by the number reported in the cell). Blank cells indicate that the statistic is not relevant (e.g., a blank cell for the number of $2km$ markets within the $2km$ or $4km$ super-market).

Table 4: Distribution of Market Size within Super-Markets

2km Super-Market						
	Number of Markets					
Rank	1	2	3	4	5	6+
1	100%	73%	66%	64%	61%	58%
2		27%	22%	19%	16%	12%
3			12%	11%	11%	7%
4				6%	7%	5%
5					4%	4%
6+						14%

4km Super-Market						
	Number of Markets					
Rank	1	2	3	4	5	6+
1	100%	72%	62%	56%	51%	44%
2		28%	24%	22%	20%	11%
3			14%	13%	14%	6%
4				9%	10%	5%
5					6%	4%
6+						30%

8km Super-Market						
	Number of Markets					
Rank	1	2	3	4	5	6+
1	100%	74%	64%	58%	58%	38%
2		26%	23%	20%	18%	11%
3			13%	13%	11%	6%
4				9%	8%	5%
5					5%	4%
6+						37%

Notes: Table reports the distribution of area share of MODIS 1km markets within $2km$, $4km$ and $8km$ super-markets. For example, in the first panel, for $2km$ super-markets that contain three MODIS 1km markets, the largest market accounts for 67% of the markets' area, the second largest market for 22%, and the smallest market accounts for 12% of area. Numbers may not sum to one because of rounding.

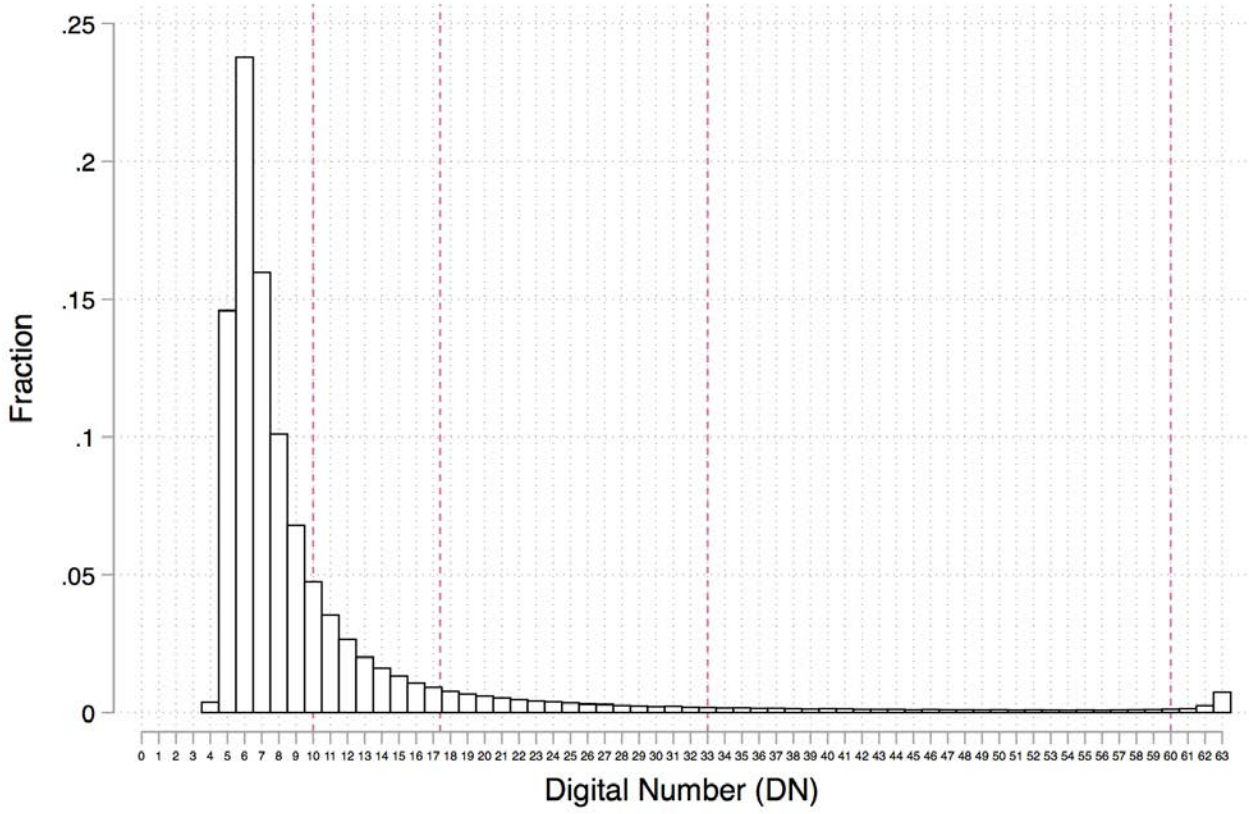
Online Appendix Tables and Figures

Table A1: Administrative Areas in India, 2011 Census

	Number	Total Population	Population Share	Mean Population	Mean Area (km^2)
Villages	640,932	833,748,852	68.9%	1,301	4.8
Towns	6,171	377,106,125	31.1%	61,109	16.6
Class 1 (>100k)	468	264,745,519	21.9%	565,696	97.6
Class 2 (50k-100k)	474	32,179,677	2.7%	67,890	20.4
Class 3 (20k-50k)	1,373	41,833,295	3.5%	30,469	14.4
Class 4 (10k-20k)	1,683	24,012,860	2.0%	14,268	9.3
Class 5 (5k-10k)	1,749	12,656,749	1.0%	7,237	5.5
Class 6 (<5k)	424	1,678,025	0.1%	3,958	4.1

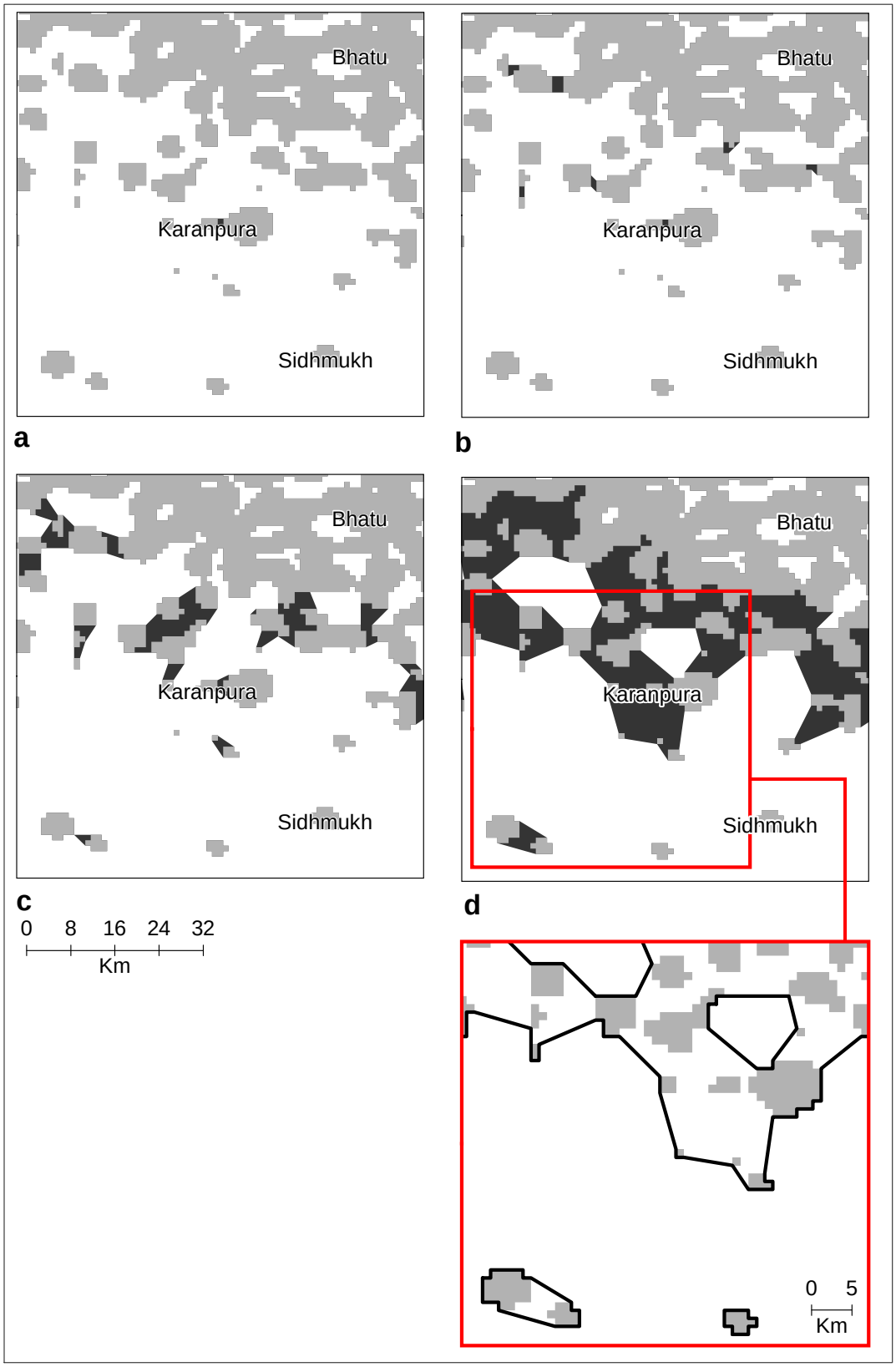
Notes: Table reports official tabulations from 2011 Census of India.

Figure A1: Density of Nighttime Lights for $1km$ Pixels, All India



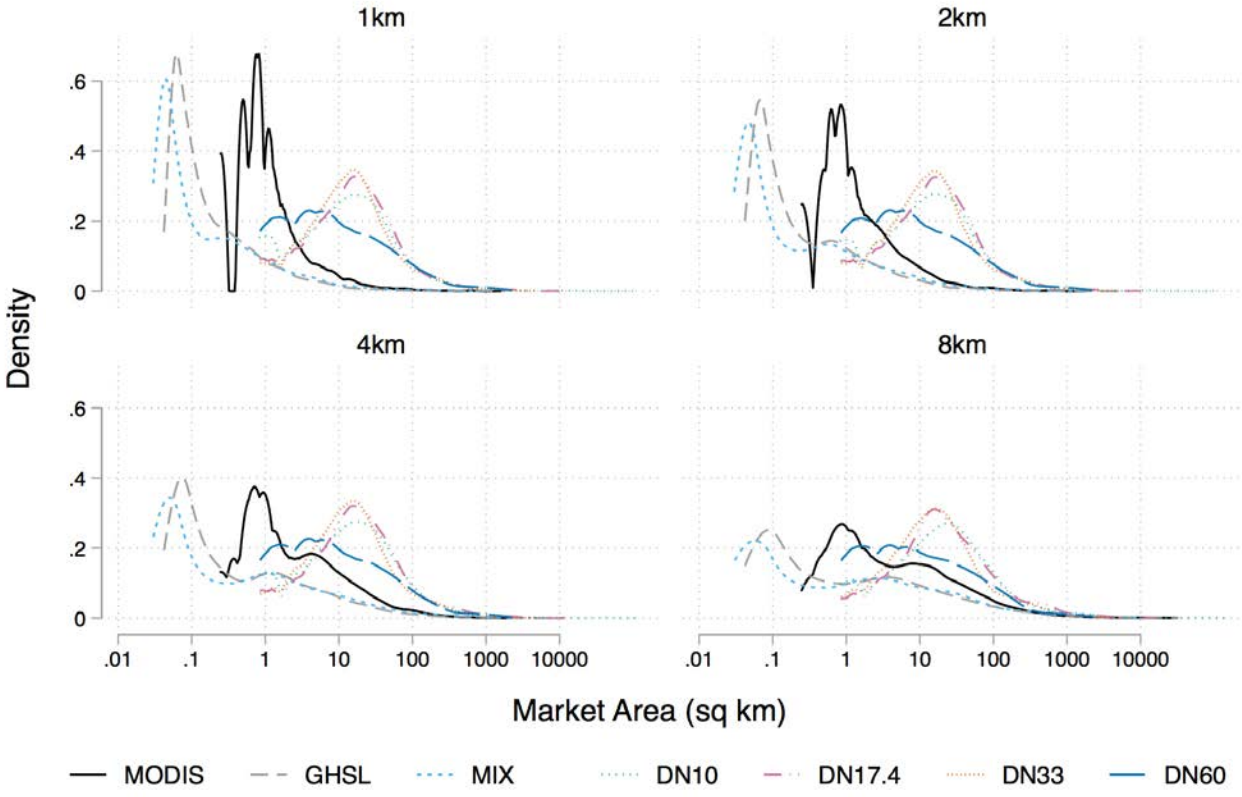
Notes: Vertical lines denote the 90th, 95th, 99th, 99.5th percentiles of DNs. Histogram formed using a 3% random sample of pixels.

Figure A2: Combining Polygons to Form Markets



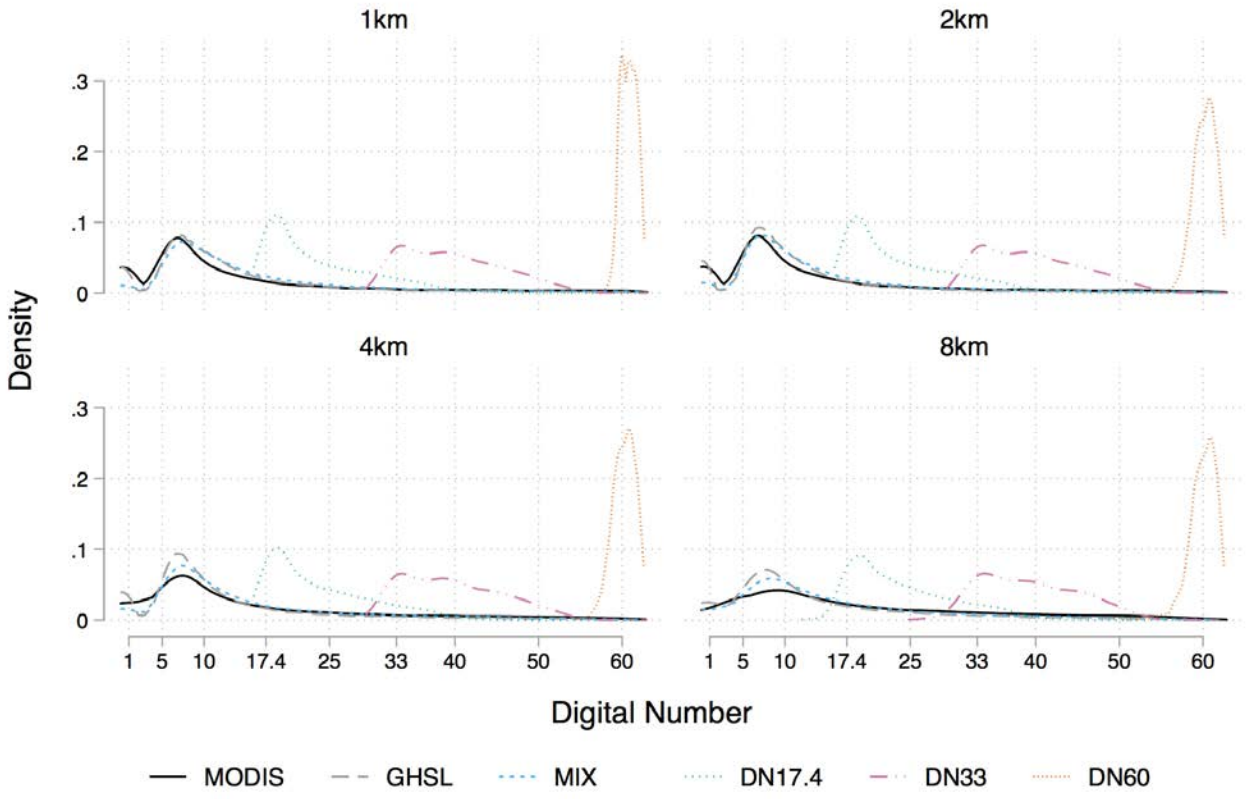
Notes: Panel (a) illustrates DN10 threshold markets. Panels (b-d) shows $2km$, $4km$ and $8km$ buffers, respectively. The last panel shows the aggregated $8km$ buffered markets.

Figure A3: Distribution of Land Area



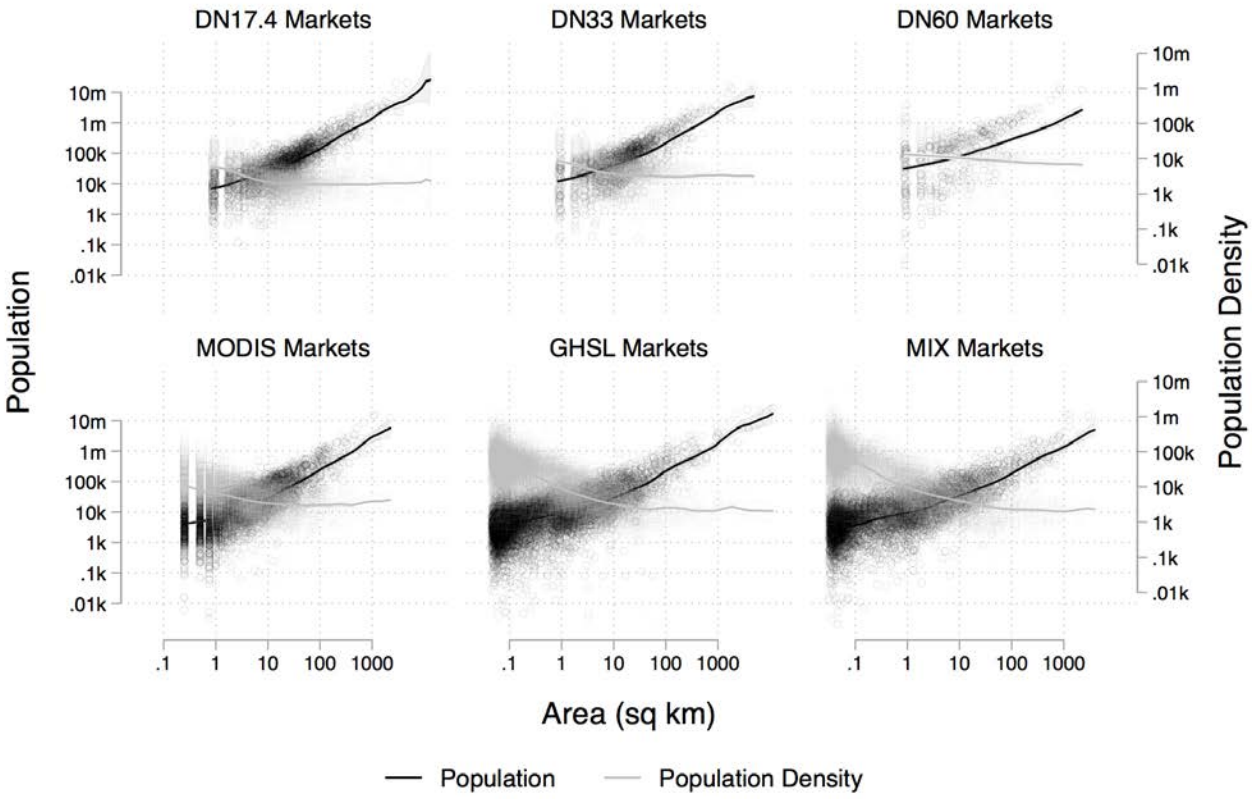
Notes: Figure reports the distribution of market land area, by market definition.

Figure A4: Distribution of Minimum Nightlight DN Values



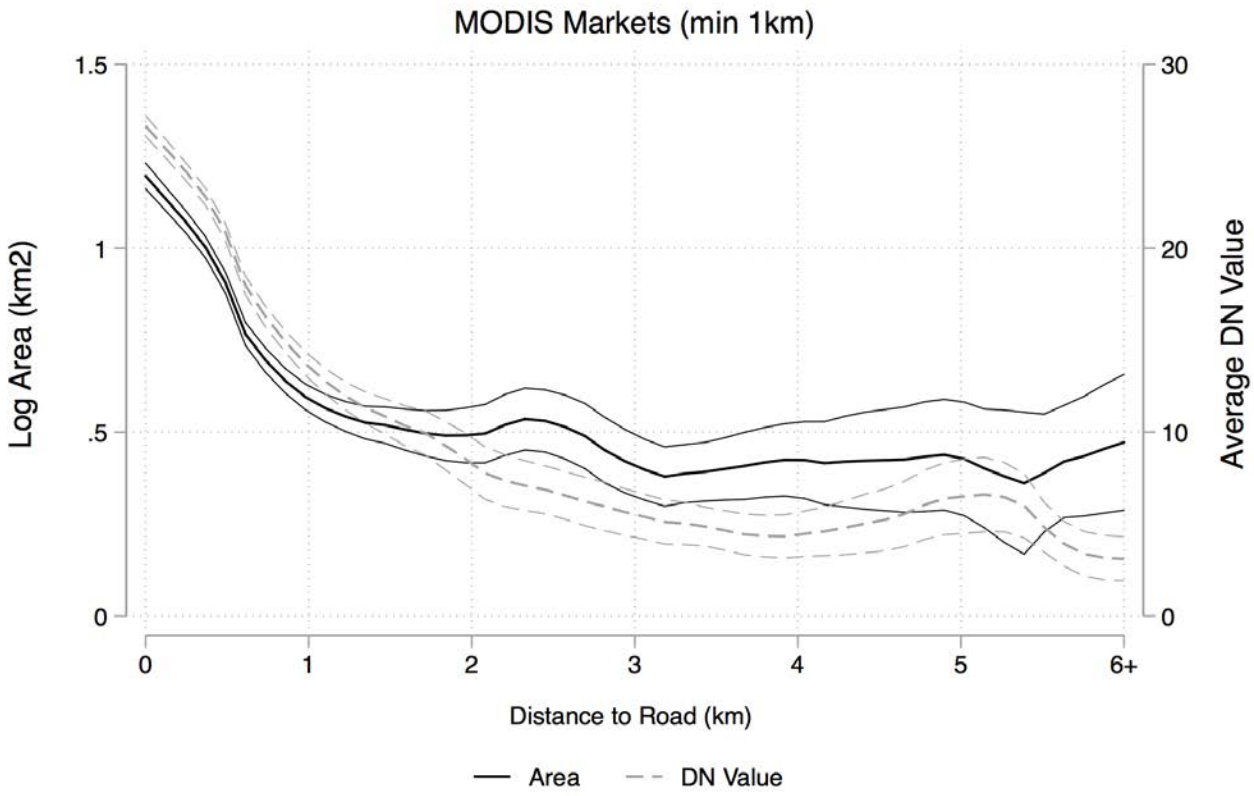
Notes: Figure reports the distribution of minimum DN values, by market definition.

Figure A5: Population versus Land Area, $4km$ Buffer



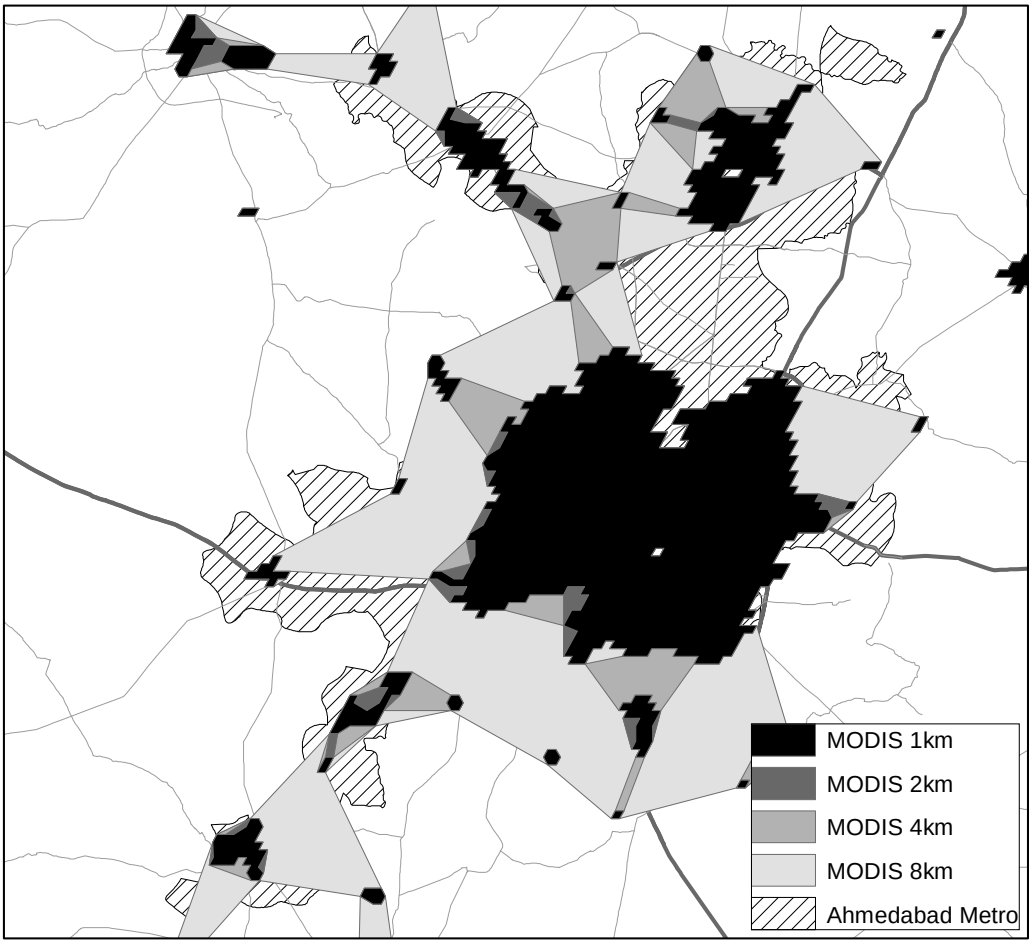
Notes: Figures reports relationship between market size, population and population density. Markets are buffered at $4km$. Population from 2011 Census.

Figure A6: Proximity to Roads, Coarser MODIS Markets ($1km^2$ minimum area)



Notes: Distance to road is the shortest distance from market centroid to a primary, secondary or tertiary road. Road data obtained from OpenStreetMaps. Figure uses MODIS markets formed using a minimum threshold of $1km$ and buffered at $1km$. Figure shows 5% and 95% confidence intervals.

Figure A7: MODIS Landcover-Based Markets within Ahmedabad Metro Area



Notes: Map shows MODIS markets in the Ahmedabad metropolitan area. The black outline is the official administrative boundary of Ahmedabad from 2011 Census. Within the administrative boundary, there are 15 $1km$, 11 $2km$, 7 $4km$ and 2 $8km$ markets.

Figure A8: MODIS Landcover-Based Markets within Ajmer Metro Area

Notes: Map shows MODIS markets in the Ajmer metropolitan area. The black outline is the official administrative boundary of Ajmer from 2011 Census. Within the administrative boundary, there are 1 $1km$, 1 $2km$, 1 $4km$ and 1 $8km$ markets.

A Aggregating Pixels to Markets

To combine clusters of highly lit pixels, we use the Aggregate Polygons function in ArcGis. This function combines polygons within a specified buffer to form larger polygons. Appendix Figure A2 illustrates the tool with lit pixels, focusing the border between Rajasthan and Haryana, two states in India. The gray areas illustrate polygons that are contiguous sets of pixels with a DN that exceeds 10. Notice that there are many unconnected polygons. Merging two polygons forms a larger polygon that contains the land area of the original two polygons plus a land bridge that connects them, whose dimension is determined by the algorithm. The larger is the distance buffer, the larger will be the land bridges that connect polygons. Figure A2a illustrates the results of implementing a $1km$ buffer; Figures A2b through A2d implement $2km$, $4km$, and $8km$ buffers, respectively. For a sub-area within the sample geographic region, Figure A2e illustrates the resulting markets when we impose the $8km$ buffer. Notice that moving from the smallest to the largest buffer collapses the number of markets in this area from more than 20 to just 3.

B Construction of the MIX Layer

This online appendix provides an overview of the builtup classification methodology developed by Goldblatt et al. (2018) for India, Mexico, and the U.S. The methodology uses DMSP-OLS nightlight data as quasi-ground truth to train a classifier for builtup land cover using Landsat 8 imagery. The basic idea is that since lights indicate the presence of human activity, we can train a classifier that uses the spectral signature of daytime images to predict the presence of humans, as indicated by lights. The challenge of using nightlights as a source of ground truth is the blooming of lights. Goldblatt et al. (2018) correct for this blooming as follows. Using their approach and imagery for 2013, we calculate the per-band median values from a standard top-of-atmosphere calibration of raw Landsat 8 scenes. These per-pixel band values are then used to construct commonly used indices to detect vegetation (the normalized difference vegetation index, NDVI), water (the normalized difference water index, NDWI), physical structures (the normalized difference built index, NDBI), and other relevant features. We use these indexes to mask out pixels that appear with high DN from the DMSP-OLS data; the assumption is that these pixels, because they are composed mostly or entirely of water or vegetation, do not contain builtup activity and appear unlit only because of blooming. We then proceed with the classification.

The steps of the methodology are as follows:

1. Designate a pixel as *builtup* if its DN exceeds a threshold. This threshold is set at the 95^{th} percentile of pixels in the training set, which is 17.4 across all India but ranges is allowed to vary across hex-cells (discussed below).
2. Re-classify a builtup pixel as *not builtup* if the Landsat index bands (NDVI, NDWI, NDBI) indicate presence of water, dense vegetation or not builtup activity (as noted above, this corrects for the blooming).

3. Use supervised machine learning to train a classifier (a random forest with 20 trees) with the adjusted builtup/not builtup binary pixels from steps 1 and 2, and the Landsat 8 median-band values and index values as inputs.
4. Use the classifier to construct the posterior probability that a pixel is builtup, and then create binary values of builtup/not builtup status based on this probability (discussed below).
5. Evaluate the accuracy of the classifier by comparing the predicted builtup status of a pixel to a ground-truth dataset that has 85,000 human-labeled pixels that were classified as builtup or not builtup.

In (3), we allow for variation in how the reflectance of India's heterogeneous land cover is associated with urbanization by partitioning the country into an equal-area hexagonal grid with hex-cells that have center-to-center distances of 1-decimal degree, and then treat each hex-cell as an independent unit of analysis. (We also train classifiers for hex-cells that have distances of 4- or 8-decimal degrees, but find that the 1-decimal degree hex-cell is most accurate.) After training the classifier separately within each hex-cell, we mosaic the resulting local classifications to map predicted builtup land cover for the entire country. In (4), we designate a pixel as builtup if its posterior probability exceeds a given threshold that is determined by the Otsu algorithm (Otsu 1979), which is a nonparametric and unsupervised method for automatic threshold selection originally developed for picture segmentation. The method uses a discriminant criterion to identify an optimal threshold that maximizes the between-class variance. We choose the threshold to maximize the variance between builtup and not-builtup classes. In (5), which compares our predicted values of builtup status with human-labeled examples, we achieve an overall accuracy rate is 84%. The accuracy rate is defined as the sum of true positives and true negatives divided by the total sample. Note that this accuracy rate exceeds the MODIS classification accuracy by 2.5% in India; see Table 6 of Goldblatt et al. (2018).