Detecting Urban Markets with Satellite Imagery: An Application to India*

Kathryn Baragwanath Vogel†
UC San Diego

Ran Goldblatt‡
New Light Technologies

Gordon Hanson§
UC San Diego & NBER

Amit K. Khandelwal¶
Columbia & NBER

First Draft: July 2018
This Draft: July 2018

Abstract

This paper proposes a methodology for defining urban markets based on economic activity detected by satellite imagery. We use nighttime lights data, whose use in economics is increasingly common, to define urban markets based on contiguous pixels that have a minimum threshold of light intensity. The coarseness of the nightlight data and the blooming effect of lights, however, create markets whose boundaries are too expansive and too smooth relative to the visual inspection of actual cities. We compare nightlight-based markets to those formed using high-resolution daytime satellite imagery, whose use in economics is less common, to detect the presence of builtup landcover. We identify an order of magnitude more markets with daytime imagery; these markets are realistically jagged in shape and reveal much more within and across-market variation in the density of economic activity. The size of landcover-based markets displays a sharp sensitivity to the proximity of paved roads that is not present in the case of nightlight-based markets. Our results suggest that daytime satellite imagery is a promising source of data for economists to study the spatial extent and distribution of economic activity.

Keywords: Satellite, Landsat, Nightlight Data, Market Access, Cities, Urbanization

JEL Classification: R1, O1, O18

*We acknowledge funding from the World Bank and from International Growth Centre (Project 89448). We thank Gilles Duranton, Somik Lall, Trevor Monroe, Rinku Murgai, Siddharth Sharma, and seminar participants at the University of Toronto and McGill University for valuable feedback.

†Department of Political Science, UCSD, 9500 Gilman Dr., La Jolla, CA 92093, *email*: kbaragwa@ucsd.edu

‡*email*: ran.goldblatt@newlighttechnologies.com

§School of Global Policy and Strategy, UCSD, 9500 Gilman Dr., La Jolla, CA 92093, *email*: gohanson@ucsd.edu

¶Columbia Business School, Uris Hall, 3022 Broadway, New York, NY 10027 *email*: ak2796@columbia.edu

1 Introduction

In the study of economic geography, cities are a standard unit of analysis. They attract labor and capital, while also facilitating investment in shared infrastructure, thereby marshaling the resources that sustain national economies. Viewing cities as the locus of economic activity is core to theoretical literature on systems of cities (e.g., Henderson 1974; Duranton and Puga 2001, Rossi-Hansberg and Wright 2007) and to work on the role of market access in shaping the spatial distribution of economic activity (e.g., Harris 1954; Krugman 1991; Redding and Venables 2004; Donaldson 2018). In empirical work in economic geography, however, data constraints often complicate studying urban markets in a manner that is commensurate with their theoretical counterparts. Common data sources, such as economic censuses, population censuses, and related surveys, tend to provide information at the level of officially designated administrative units that map imperfectly to urban markets, ranging, for example, from Indian districts or U.S. counties, which may comprise many distinct urban markets (e.g., Hanson 2005; Donaldson and Hornbeck 2016; Ghani et al. 2014), to jurisdictionally defined towns and urban places, whose connection to urban areas may be difficult to decipher (Eeckhout 2004, Rozenfeld et al. 2011). In many developing countries, these data sources are collected at low temporal frequencies and are often difficult to obtain.

Remotely-sensed data, including satellite imagery, offer a potential solution to the mismatch between theory and traditional data sources (Donaldson and Storeygard 2016). Economists have used satellite data on the intensity of light emitted at night to study national economic growth (Henderson et al. 2012), regional economic development (Gennaioli et al. 2013), and the global distribution of economic activity for $1km^2$ grid cells (Henderson et al. 2018), among a rapidly growing set of topics. Daytime satellite imagery, whose use in the study of urban sprawl was pioneered by Burchfield et al. (2006), is available at even higher spatial and temporal resolutions, down to $30m$ at a biweekly frequency for data going back to the late 1990s and down to less than $1m$ resolution at a daily frequency for imagery from recently launched proprietary satellites. After long suffering from a paucity of geographically disaggregated information on economic activity, economists are now awash in data at previously unimaginable spatial and temporal resolutions.[1] The challenge is no longer how to shoehorn data on administrative units into economic models, but rather how to translate high-dimensional data into a form that generates insights for research and policy.

This paper develops a new methodology to delineate urban markets using satellite data. Our goals are to produce a flexible and globally scalable measure of urban markets that matches the conceptional definition of these markets in standard models of economic geography, and to examine how markets detected from nightlight satellite imagery compare to those detected from daytime high-resolution satellite data. Because our measures are algorithmic, we are not exposed to the vagaries of administrative boundaries across space. To motivate our approach and to highlight the challenges that we confront, consider the inherent messiness of the spatial structure of urban areas.

[1] The Landsat 7 program, for instance, has hundreds of terabytes of data and adds 260 gigabytes of a data to the archive daily (Wulder et al. 2016).

Seen from above, cities are clusters of variably-sized settlements connected by a complex web of transportation corridors. These networks appear to have a fractal dimension: the irregularity and the interconnectedness of their component clusters is maintained as one increases the spatial resolution. At a coarse resolution, the Hyderabad metropolitan area of India has an amoeba shape, with pointy arms that extend north, northwest, southwest, and southeast, along major highways. As we zoom in, the amoebas multiply. Hyderabad and Secunderabad appear as the larger masses, Ghatkesar and Kukatpally are larger satellites around these masses, and more amoeba-like satellites appear as we zoom in further. For empirical analysis, do we think of Hyderabad—a city that has 6.8 million people spread over $650 km^2$— as a single market or as multiple markets? How does the spatial definition of markets affect the empirical application of standard economic models?

We define a market as a set of contiguous, or near contiguous, pixels that contain economic activity. Satellite images of nighttime lights are one source of data that indicates the presence of economically active agents (Henderson et al. 2012). Operationalizing this definition using nightlight data requires choosing a minimum threshold of light intensity for the contiguous pixels. Following Rozenfeld et al. (2011), we experiment with buffers that combine contiguous sets of pixels if they lie within a radius of $1km$, $2km$, $4km$ or $8km$. Having to choose a nightlight threshold to define a nightlight-based market immediately reveals a tradeoff: while a strict threshold only captures major urban agglomerations, lowering the threshold to include smaller cities comes at the expense of exploding the size of larger cities that have satellite towns. This tradeoff is in part a consequence of the blooming effect of light, which tends to produce cities whose boundaries are too expansive and too smooth relative to the haphazard shape of 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 much finer resolution than nighttime light data but require image classification techniques to detect the spectral signature of a pixel. We explore data on builtup landcover from two publicly available layers: one based on imagery from the MODIS satellite sensor and constructed by Channan et al. (2014), and another, the Global Human Settlements Layer, based on Landsat satellite imagery and developed by Pesaresi et al. (2015). We also examine a recently developed layer of builtup landcover for India by Goldblatt et al. (2018). Using an algorithm that defines markets as a set of contiguous, or near contiguous, pixels of builtup landcover, we find that these landcover-based markets, irrespective of the data source, are realistically jagged in shape and reveal substantially more variation in the density of economic activity when compared to nightlight-based markets. We then combine the two data sources to measure the extensive and intensive margins of economic activity across markets. This exercise reveals that daytime imagery is well-suited for defining the extent of market areas, and that nightlight imagery is useful for capturing the intensity of activity within these market boundaries. In other words, daytime imagery can be used to accurately measure the boundary of a market, while the economic activity that the market generates can be measured by the area's nightlight intensity.

Our approach explores alternative definitions of urban markets based on different data sources

and different buffering thresholds, while remaining agnostic as to the underlying economic determinants of market size. Some definitions, particularly those with coarser distance buffers for grouping pixel clusters, may correspond to markets within which internal trade costs are low such that they self-contain retail transactions, personal services, and other commerce that economists typically consider to be non-traded activities (Redding 2016). By contrast, definitions with finer distance buffers may approximate commuting zones within which individuals live and work (Duranton 2015). Researchers who follow our method may choose different buffers to define markets based on the conceptual foundations of cities that underly their applications.

To preview our findings, we detect an order of magnitude more urban markets through daytime satellite imagery than with nightlight satellite imagery. Using the definition of a market that buffers clusters of contiguous pixels at $1km$, we observe 13,314 markets from MODIS data compared to 1,669 markets from a nightlight threshold of 33 (the 98th percentile in India) or 469 markets from a nightlight threshold of 60 (the 99.5th percentile, where the maximum nightlight value is 63). Using the layer for India developed by Goldblatt et al. (2018), we detect 17,304 markets, while the Global Human Settlements Layer (GHSL) detects 26,202 markets.[2] Daytime satellite imagery reveals remote towns that are distant from India's urban centers. For instance, we observe some landcover-based markets that have an average nightlight intensity of just 5. This suggests that we are able to capture many rural parts of India that do not have reliable access to electricity.[3] 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. Landcover-based markets are therefore able to capture small cities and towns in India while preserving the spatial distribution of activity of the largest cities. Moreover, within large cities, we can detect many distinct neighborhoods, which facilitates evaluating the impacts, for instance, of local infrastructure within polycentric cities (Duranton and Puga 2015).

Next, we document some intriguing additional characteristics of landcover-based markets. As compared to nightlight-based markets, these markets more closely follow a power law in land area (and for MODIS markets buffered at $1km$, we observe an area-rank coefficient that is very close to Zipf's law). This is consistent with Rozenfeld et al. (2011) who document a Zipf's law for market land areas in the U.S. and Great Britain. We also observe a positive correlation between the area of landcover-based markets and measures of population, population density, and average nightlight values.[4] These correlations are important for assessing a limitation of daytime satellite data. While these data are suitable for measuring the extent of a market, they cannot reveal the intensity of economic activity within a market. The positive correlations we uncover suggest that larger markets in terms of land area also have higher levels of economic activity. For example, a MODIS market at the $25th$ percentile of the land-area distribution has a nightlight intensity of 7.6

[2] Section 2 discusses these data sources in detail.

[3] According to the World Development Indicators, 1.121 billion people in the world did not have access to electricity in 2010, of which 392 million were located in rural South Asia.

[4] The population data come from WorldPop (http://www.worldpop.org.uk/).

compared to 16.8 for a market at the $75th$ percentile of land area. Using the nightlight-GDP elasticity of 0.3 from Henderson et al. (2012), the larger market would have a GDP that is 21% greater.

Finally, we use landcover-based markets to examine spatial variation in market access. A key determinant of market access is proximity to transportation networks (e.g., Donaldson and Hornbeck 2016). For each market definition, we compute the distance between the centroid of a market and the closest paved road. Landcover-based markets exhibit a sharp negative elasticity in size and nightlight intensity with respect to nearest-road distance: a landcover-based market whose nearest paved road is just $2km$ away is roughly half as large as one with a road that bisects the market centroid. For nighlight-based markets, we are unable to detect an association between market size and nearest-road distance over short distances. This could be due either to the relatively coarse resolution of nightlight data or to the blooming of lights. Our results suggest that blooming is the larger concern. When we aggregate daytime data to the spatial resolution of the nightlight data and reform coarser landcover-based markets, these markets still reveal a sharp negative area and nightlight elasticity with respect to road proximity. It thus may be possible to estimate micro-spatial impacts of transportation infrastructure on market size and economic intensity through landcover-based markets that would not be identifiable using markets based solely on nightlight data, especially in regards to "last-mile" challenges in delivering goods and services. We also demonstrate that a non-trivial portion of a market's total market access, following Donaldson and Hornbeck (2016), is determined by other distinct "sub-markets" that lie within a larger buffered "super-market". This suggests that granular impacts of intra-city infrastructure, such as ring roads or metro rail, are detectable from daytime imagery.

Our results contribute to recent literature on methods to delineate markets that do not rely on administrative boundaries. The paper most closely related to ours is Rozenfeld et al. (2011), which builds an algorithm based on near-contiguous populated areas in Great Britain ($200m$ grid cells) and the U.S. (FIPS codes). The advantage of their data is that they are able to measure total area and total population in markets constructed "from the ground up." However, they rely on high-quality micro data, which are unavailable for most countries, particularly those in the developing world. Our paper also 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, and Burchfield et al. (2006), who use contiguous pixels to measure sprawl in the U.S. based on Landsat satellite imagery from 1976-1992. Recent work by Duranton (2015) proposes alternative an algorithm to construct markets based on commuting patterns for Colombia. In concurrent work, Davis et al. (2018) also use clusters of pixels above nightlight thresholds (what we refer to as nightlight-based markets) to construct metro areas in Brazil, China and India. While our interest is in comparing the use of nightlight imagery with daytime imagery and evaluating their potential for studying market access, their interest lies in examining the distribution of skills across space within those countries.

In section 2, we present alternative methods to detect markets from nightlight and daytime satel-

lite imagery. In section 3, we compare the properties of nightlight-based markets and landcover-based markets. In section 4, we examine how landcover-based markets can be used to evaluate market access. And in section 5, we conclude.

2 Algorithmic Approach to Detect Urban Markets

We define markets using two sources of remotely-sensed data: (1) the intensity of light emitted at night as captured by DMSP-OLS (nighttime lights data), and (2) classifications for builtup land-cover based on three daytime satellite imagery products. In this section, we describe the data sources and algorithms 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 nighttime from different sources on Earth, such as city lights, auroras, gas flares, and fires. Pixels in DMSP-OLS have a resolution of 30 arc seconds, or approximately $1km \times 1km$. For each pixel, the digital number (DN) of calibrated light intensity ranges from 0 to 63, which we refer to as the nightlight value or intensity.[5] 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.[6] 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 and map the spatial distribution of lit pixels according to the satellite data in that year.

There are well-known limitations to DMSP-OLS data (e.g., Donaldson and Storeygard 2016). 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 the combined effect of several factors related to DMSP-OLS data: (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-size sized ellipse, but its centroid is shocked in a random compass direction (Abrahams

[5] Tuttle et al. (2014) place 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 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.

et al. 2018); and (3) on-board data management, where the 1970s technology on board the satellites causes top-censoring of inputs.[7] The highest possible DN of DMSP-OLS pixels is 63, and because of this saturation, it is often impossible to differentiate between medium-density cities and high-density cities.[8] In our setting, saturation is not an issue because we measure the extent of markets through lower bounds of light intensity. However, blooming is potentially 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.*

To operationalize this definition of a market based on nightlight data, three choices are required: 1) what is the minimum number of pixels that constitute a market; 2) what parameters should govern "near contiguity"; and 3) what minimum DN should be used.

As mentioned above, the DMSP-OLS sensor has a $1km$ resolution. We set the minimum number of pixels to form a market at 1 pixel.[9]

To determine the minimum DN thresholds for our market definition, we examine the distribution of DNs across pixels in India for 2013 in Appendix 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,[10] 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. Motivated by these patterns, we set the following alternative thresholds for a pixel to be highly lit: 10 (90^{th} percentile), 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

[7] Fine pixel values are summed together in consecutive 5 × 5 blocks to form an image of 2.8 km × 2.8 km coarse pixels. To further economize data management, the smallest two bits of each coarse pixel's value are dropped, meaning that all coarse pixel values are divided by 4, integized, and top-censored at 63, producing the 6-bit quantization familiar to users of these data (Abrahams et al. 2018).

[8] See Henderson et al. (2018) for analysis that uses a radiance-calibrated version of the nightlight data (Elvidge et al. 1999), which alleviates the saturation effect. These data are available for only a subset of recent years.

[9] 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.

[10] 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.

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 $8km$.[11] This expansive set of distance buffers allows us to check the sensitivity of our method to the distance threshold for combining adjoining clusters of highly lit pixels. Different buffers have different interpretations. In this paper, we are agnostic as to the choice a benchmark buffer. For some applications, researchers may be interested the analyzing the responsiveness of $1km$ markets, while in other applications they may want to smooth over this level of granularity.

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 Harayana, 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.

2.2 Detecting Markets from High-Resolution Daytime Imagery

Daytime imagery offers an alternative approach 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 classification of the landcover. In recent years, there has been substantial 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 two publicly available sources of landcover classification from daytime imagery for our analysis, one which is based on imagery from MODIS (Channan et al. 2014)[12] and another of which is the Global Human Settlements Layer (Pesaresi et al. 2015), to which we refer as GHSL, which is based on imagery from Landsat 7 and Landsat 8.[13] Both layers are derived from applying supervised machine learning to train a classifier on the builtup status of a pixel with the spectral properties of the associated images as inputs. MODIS uses a supervised machine learning method which

[11] 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.

[12] The Moderate Resolution Imaging Spectroradiometer (MODIS) was launched by NASA in 1999. It has a temporal resolution of one to two days and 36 spectral bands, which range in resolution from $250m$ to $1km$.

[13] 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, consists of nine spectral bands with a spatial resolution of $30m$ at a temporal frequency of 16 days.

takes advantage of a global database of training sites extracted from high-resolution imagery. We use the University of Maryland yearly classification scheme (Channan et al. 2014), which is available from 2001 forward with a 500m resolution. We use MODIS data for 2013, and take the Urban and Builtup pixels (classification 13) to indicate builtup landcover. The GHSL layer combines satellite data from the Global Land Surveys datasets (GLS1975, GLS1990, GLS2000) and Landsat 8 (Pesaresi et al. 2015) to determine builtup grids at a 38m spatial resolution. We use their Built Up Confidence Grid which aggregates builtup data in 2014, and classify pixels where the confidence of builtup is higher than 50% to identify builtup pixels.

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, is based on using DMSP-OLS nightlight data as quasi-ground truth to train a classifier for builtup land cover with daytime satellite imagery as inputs.[14] This method has a higher accuracy rate than the MODIS layer, but at present, we have only constructed it for India, U.S. and Mexico. See Appendix A for a description of the methodology and the published paper for details.

Using these three different layers that classify builtup landcover—MODIS, GHSL, and MIX—we adopt 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 they contain builtup landcover.*

For MODIS markets, we impose a minimum number of pixels for a market to be 1 ($0.063km^2$). For GHSL and MIX, the minimum number of pixels is set to 40 ($0.058km^2$ and $0.036km^2$, respectively). Choosing a minimum pixel size of 1 for GHSL and MIX would be a fairly extreme choice given the granularity of these data (and is computationally very expensive to process); the choice of 40 takes advantage of the granular data to detect small clusters of pixels while not resulting in artificially small markets which would rarely constitute areas with well-defined internal trade or self-contained commuting. Clusters of builtup pixels are aggregated in an analogous way as described above using the Aggregate Polygons function in ArcGis.[15]

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 and for each of the four buffers, we plot the markets detected around three cities in India: Delhi (population in 2011 of 19 million), Ahmedabad (2011 population of 5.5 million), and

[14] In contrast to previous methods that rely on relatively small samples of ground truth to train and validate classifiers (e.g., Goldblatt et al. 2016), the use of nightlights as an input into the classifier allows one to scale image classification across countries and regions without requiring manually classified ground-truth data, which are expensive and time consuming to collect.

[15] For example, if two clusters of 45 pixels are separated by, say, 1.5km of non-builtup pixels, they would form two markets under the 1km buffer and one market under the 2km buffer.

Ajmer (2011 population of 0.5 million) in Figures 1 to 3.[16] We overlay road networks from OpenStreetMaps in 2018 to give a visual sense of how transportation networks may tailor the shape of markets. Panels (a) to (d), in the first row, display results for MIX-based markets, while panels (e) to (t), in the second through fifth rows, display results for nightlight-based markets.

Consider first nightlight-based markets. Together, we have 16 alternative nightlight-based market definitions in panels (e) to (t) of Figures 1 to 3. The maps clearly 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, such as Meerut (Uttar Pradesh, 1.3 million), Rohtak (Haryana, 0.4 million) and Bhiwadi (Rajasthan, 0.1 million). The blob itself is 12,555 km^2, which is close to the size of the U.S. state of Connecticut. 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 about 2 hours to the northeast by car from 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, one can see that activity appears to coalesce along the artery. This is problematic as these lights are likely capturing street lights and car lights along the road rather than clusters of economic activity.

To compare with landcover-based markets, examine the top rows of Figures 1 to 3, which show markets using the MIX layer. Results for MODIS and GHSL layers are similar. In stark contrast to the nightlight-based definition in the bottom four rows, landcover-based markets are jagged in shape and display substantial within-market variation in the density of economic activity.[17] 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 urbanized. These clusters include bodies of water, parks, and undeveloped land. 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 the ex-

[16] Population numbers are taken from the 2011 Census.

[17] This is confirmed by the higher perimeter to area ratio of landcover-based markets compared to nightlight-based markets.

tent of urbanization inside market boundaries appear to be overly smooth. Excessive smoothness may distort measures of the ease of distributing goods inside markets if such measures were to be based on variation in the density of economic activity within market boundaries. 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, which could have the interpretation of integrating markets through the trading of goods and services, while remote towns remain visible.

Visual inspection makes apparent 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 is 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.

3 Comparing Nightlight- and Landcover-Based Markets

Having constructed markets for India using either pixel clusters whose light reflection exceeds a given threshold (nightlight-based markets) or pixel clusters indicative of builtup landcover (landcover-based markets), we compare market characteristics across these definitions. We examine the number markets detected, market land area, the applicability of Zipf's law, and the intensity of economic activity (as captured by population and nightlight intensity) within market boundaries.

3.1 Detecting the Number of Urban Markets

Whereas theories of urban structure have much to say about the size distribution of cities (e.g., Duranton 2007; Rossi-Hansberg and Wright 2007), they have less to say about the number of urban places that populate a distribution. The formation of such markets is constrained by the fixed costs of building transportation networks and other infrastructure, among other factors (e.g., Henderson and Venables 2009). Because some types of infrastructure—such as major highways and power plants—can be shared across cities, they may play as much of a role in determining population density at the state or provincial level (e.g., how many people live in Gujarat) as in dictating how the population is distributed across space within these areas (e.g., do people in Gujarat live in 10 urban markets or in 100 markets). These within-region population distributions may be shaped, in turn, by scale economies in retail distribution and establishment-level production. Such scale economies, however, are not rigidly determinative of urban structure (Lucas and Rossi-Hansberg

2003). When we contemplate the number of urban markets we expect to detect in a country, we enter with few expectations beyond the size distribution of these markets being heavy tailed.

The top row of Figure 4 illustrates the number of markets detected through nighttime lights. As expected, 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 5,334 DN 10 markets, 3,275 DN 17.4 markets, 1,669 DN 33 markets, and 469 DN 60 markets. The two higher DN thresholds exhibit little variation in the number of markets across buffers.

For context, we compare the number of nightlight-based markets detected through our algorithm with official enumerations from the Census of India. Appendix 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 377 million people. 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. This official definition therefore combines administrative boundaries with constraints regarding population size, population density, and type of economic activity. Of the 6,171 towns, 468 are considered "Class 1" cities with more than 100,000 inhabitants; these are the largest urban centers of India and correspond closely to the number of markets detected with a DN threshold of 60 (at any distance buffer). There are 1,847 Class 1, 2 and 3 towns—localities with at least 20,000 inhabitants—a number roughly similar to the number of DN 33 markets at a $1km$ buffer. There are a further 1,683 towns with 10,000 to 20,000 inhabitants. Comparing Figure 4 and Appendix Table A1, we see that only DN 17.4 markets at a $1km$ buffer (roughly) matches the number of officially recognized Indian cities and towns with more than 10,000 residents. However, we demonstrate below that these DN 17.4 markets have some unappealing properties.

The bottom row of Figure 4 reports the number of markets detected from daytime imagery. While the numbers vary across the three different 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 13,314 at distance buffer of $1km$ to 3,469 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. 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. This discrepancy suggests that landcover-based markets capture smaller areas that tend to lie within the boundaries of officially designated metropolitan regions. That is, daytime imagery can detect "sub-markets" that lie within "super-markets", a feature that we explore in more detail below. When it comes to the analysis, for instance, of how improved transportation infrastructure affects market access, it may be useful to distinguish between investments in inter-city highways or railways, which may have their primary effects on shifting

activity between metropolitan areas, and the portions of these investments that affect intra-city transportation, which may differentially change market access within cities. Whereas landcover-based markets are amenable to both levels of analysis, nightlight-based markets appear to be suited only to studying the inter-urban distribution of economic activity.

3.2 Land area

Canonical theories of city formation posit that urban areas have a monocentric or polycentric structure (Duranton and Puga 2015). Actual cities are rarely so well ordered. Urban sprawl occurs unevenly at city boundaries. As cities grow outward, there often remains undeveloped land within city limits. Some of this land may be left undeveloped permanently in the form of parks, waterways, or other open spaces. Other land may remain undeveloped temporarily as owners wait for the right opportunity to come along, owners with conflicting claims adjudicate disputes over land titles, or governments determine how to utilize land that they own or control. When defining the size of cities in terms of area, we are interested both in the land area that lies within the outer envelope of urban development and in the share of land within this envelope that is builtup. Because of the blooming effect of nightlights, nightlight-based markets may detect urban areas that cover large tracts of land when compared to landcover-based markets. The latter, by design, allow for the exclusion of undeveloped land within cities. We now proceed to compare total land area across our urban market definitions.

Figure 5 reports average market size in terms of land area by market definition. Consider nightlight-based markets, first. For a DN threshold of 17.4, 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 $43.7km^2$ at a DN threshold of 60. For landcover-based markets, illustrated in the bottom panel of Figure 5, the average market sizes are much smaller. At a $1km$ buffer, MODIS markets are $2.6km^2$, while the average size of GHSL and MIX markets are $1.4km^2$ and $1.9km^2$, respectively. The smaller market sizes are a result both of the the granularity of the daytime imagery and of the exclusion of non-builtup land area (e.g., due to blooming). At a $4km$ buffer, the size of MODIS, GHSL and MIX landcover-based markets rise to $8.3km^2$, $10.9km^2$, and $12.1km^2$.

Appendix Figure A3 provides a comprehensive comparison by plotting the distribution of market sizes by definition. The mode of each distribution effectively reveals the minimum number of pixels used to define a market. MIX markets, which are constructed using $30m$ resolution images and a minimum of 40 pixels, have the smallest mode, followed by GHSL markets (based on imagery with a $38m$ resolution and a minimum of 40 pixels), then by MODIS markets (based on a resolution of $250m$ and a minimum of 1 pixel), and finally by nightlight-based markets (with a $1km$ resolution and a minimum of 1 pixel). 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.

As a further means of differentiating among market definitions, it is useful to compare maximum market sizes. The maximum area of MODIS, GHSL and MIX markets at a $1km$ buffer are 1,186km^2, 1,842km^2, and 1,097km^2, respectively. By contrast, the maximum sizes of nightlight-based markets at a $1km$ buffer across the DN 17.4, DN 33, and DN 60 thresholds are quite large: 9,977km^2, 4,681km^2, and 2,223km^2, respectively. For context, the tri-state land area of metropolitan New York City, which comprises many clusters of economic activity, is 11,642km^2. It would thus appear that nightlight-based markets may be too expansive in the land area that they comprise.

These comparisons reinforce 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 in size. This tradeoff is not present in the construction of landcover-based markets. Figure 6 reinforces this perception by plotting the distribution of average nightlight values within market boundaries. For DN thresholds of 10, 17.4, 30 and 60, the densities are left censored at the DN threshold, by construction. (Note that because of buffering these markets do capture pixels below their respective thresholds, which is naturally 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.[18] Landcover-based markets, because they are not subject to a blooming effect, span a relatively wide range of intensity in economic activity (as captured by nightlight intensity in these markets).

3.3 Zipf's Law for Market Land Areas

Urban economists have long been interested in the size distribution of cities. Understanding if city size follows Zipf's Law—a prime focus of the literature—is important for assessing the dynamics of urban growth and whether these dynamics are consistent with a random growth process or some other growth process that generates a stable cross-section distribution of city population (Duranton and Puga 2014). 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 naturally requires taking a stand on the urban expanse of cities, and with it, confronting the thorny issues of which data sources to use, and how to assess the quality of these sources and the accuracy of their implied methods for designating administrative boundaries. The motivation for the algorithmic approach developed by Rozenfeld et al. (2011) is to construct systematically the extent of urban markets, without having to rely on seemingly arbitrary boundaries, and then to test for the presence of Zipf's Law using cities whose boundaries are justified based

[18] Recall from Figure A1 that a sizable fraction of India's pixels have DNs below 10.

on economic fundamentals. In that paper, Zipf's Law for population holds quite well for the U.S. and Great Britain.[19] They additionally show that the distribution of city land areas approximately obeys Zipf's Law.[20]

We examine the emergence of a power law in the distribution of land areas for our market definitions. Figure 7 plots the log of market rank minus 0.5, based on land area, against the log of land area.[21] 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 0.87 to 0.91). While we do not find evidence of Zipf's law (the slopes, for the most part, differ from 1), landcover-based markets more closely follow the log-linear relationship dictated by a power law. The figure also reveals that for nightlight-based markets the 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 sharp contrast, the linear slopes of the area-rank plots for landcover-based markets flatten out as the buffer size increases. Figure 7 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 $1km$ buffered markets using the MODIS data source, the slope of the line is quite close to -1, suggesting that Zipf's law emerges.

3.4 Economic Activity within Markets

Our results so far illustrate that the granularity of landcover-based markets allows for more accurate measurement of market land area, when compared to nightlight-based markets. However, the daytime satellite data used to construct these markets do not provide much information on the intensity of economic activity within markets. This is because in the landcover layers the pixels record only whether or not there is a man-made impervious structure. One would need additional information on population and (or) income to predict economic activity based on the underlying spectral signatures of those images.[22] In this section, we examine the correlations between measures of economic activity and market area. Since we are confident in measuring the land area of a market, the strength of these correlations provides some assurance that land area is a reasonable proxy for both the extent of cities and the intensity of economic activity of cities.

One approach to measure economic activity within our market boundaries is to overlay publicly available layers of gridded population data, such as WorldPop.[23] Because this procedure is po-

[19] Using population counts based on administrative boundaries, Gabaix (1999) finds that the fit is near perfect for 135 U.S. MSAs, while Chauvin et al. (2016) find that Zipf's Law roughly holds in the U.S. and Brazil but not in India or China.

[20] They note that a model with Cobb-Douglas preferences for goods and housing along with a random growth process can generate Zipf's law for both population and land area.

[21] Gabaix and Ibragimov (2011) explain that this adjustment improves finite-sample properties when estimating power law exponents.

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

[23] See http://www.worldpop.org.uk/.

tentially fraught with measurement errors, we approach it with some trepidation. The underlying population data, which are collected at lower temporal frequencies and spatial resolutions than the satellite imagery, are combined with these data to generate populate estimates at fine spatial resolutions. Moreover, different prediction algorithms are used for different regions of the world. Nevertheless, these gridded population data are useful as a rough check to ensure that our market definitions are capturing places where people do, in fact, live. WorldPop tabulates population projections at $100m$ gridded intervals throughout the globe, and we overlay the 2011 layer with our market definitions. In order to assign population values to markets smaller than $100m^2$, we divide the population value of each $100m$ grid by $(100^2/30^2)$, and attach this value to each $30m$ grid. This means we assign values to the smaller $30m$ grids assuming a uniform distribution of the population inside the grids. Once each $30m$ grid has a population value, we calculate the total population count inside each market by summing the value for each grid that lies inside the boundaries of the market. For grids that do not fully lie inside the market boundaries, we include the value of such grids if more than 50% of the grid lies inside the boundaries.

Figure 8 reports the correlation between population and area for each market definition at the $1km$ buffer. For each market definition, there is a positive correlation between the area of the market and its total population. This is the case for both nightlight- and landcover-based markets. We also examine population density, defined as population divided by land area, as a measure of economic activity. Again using $1km$ buffered markets, Figure 9 shows a positive relationship between population density and area—though with greater dispersion around the regression line than for the plot in Figure 8. Larger markets in India appear to be more densely settled. If density is a proxy for economic activity (Ahlfeldt et al. 2015), this suggests that the area of landcover-based markets is a reasonable proxy for economic activity.

Previous work by Henderson et al. (2012) and Henderson et al. (2018) have demonstrated that nightlight intensity can proxy for national or regional GDP. Inspired by this work, we can analogously compare the average DN 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 landcover-based markets appear to delineate accurately the boundary of markets, we can simply compute the average DN within those boundaries. Figure 10 reports the relationship between the average DN and the land area of a market. For each of the landcover-based markets (at a $1km$ buffer), larger markets are associated with higher mean DNs. The change in DNs across market size is quite sharp. For example, a MODIS market at the $25th$ percentile of the area distribution has a mean nightlight intensity of 7.6 compared to a value of 16.8 at the $75th$ 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 21% between markets that span the interquartile range of land area. Figure 10 thus suggests that the combination of daytime and nightlight imagery may be useful for characterizing both the extensive *and* intensive margins of markets, a finding we exploit further below.

Figure 10 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 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 indicative of 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.

4 Roads and Market Access

In this section, we demonstrate the potential use of high-resolution daytime satellite imagery to evaluate market access. Transport infrastructure is an enormous expenditure for many countries. Daytime satellite imagery could become valuable source to assess the impacts of large-scale transport investments on the spatial distribution of economic activity. Their granularity allows policymakers to observe impacts both within markets (e.g., sub-markets or neighborhoods within a larger market) and across markets, and at high temporal frequencies (important for policymakers loathe to wait years to evaluate the returns to public infrastructure investments). There has been a burst of recent empirical work on the economic impacts of transportation infrastructure (Redding and Turner 2015), with the standard geographic unit of analysis being a county, municipality, or district. Satellite-based measures of markets allow such analysis to be extended to much higher spatial resolutions. While a complete analysis is outside the scope of this paper, we provide some descriptive relationships that demonstrate the usefulness of these data.

4.1 The Elasticity of Market Land Area with Respect to Road Proximity

We begin by plotting the distance between the centroids of $1km$ buffered markets to the nearest road. The road data for major and minor roads are taken from 2018 network on OpenStreetMaps.[24] 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. Absent digitized historical road maps for India, we are left with this imperfect match.

Figure 11 shows fraction of markets, for alternative market definitions, across bins defined according to proximity to the nearest paved road. As one might expect, the centroid of the vast majority of nightlight-based markets are very close to a road—within $1km$ or $2km$, in fact. The same is the case for landcover-based markets: 89.3% of MODIS markets lie within $2km$ of a road. Since we believe that most urban markets would be connected to a road of some kind, this regularity provides some validation for using daytime satellite imagery to classify markets (i.e., finding clusters that are in fact builtup pixels).

[24] We use the OpenStreetMaps road classifications. The major roads (511x) include motorways, freeways, trunk, primary, secondary and tertiary roads. We additionally include two minor road classifications: smaller local roads (5121) and roads in residential areas (5112).

Figure 12 plots the relationship between market land area and distance to the nearest paved road. This figure neatly illustrates the 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 50% smaller in land area. Such a large difference in size is not detectable using nightlight-based markets: for markets based on one of the three higher DN thresholds, the elasticity of size with respect to distance to a road is an imprecisely estimated zero. At a DN threshold of 10, we see a negative relationship between market land area and road proximity emerge only after moving to distances at least $4km$ from the nearest road.

Figure 13 repeats the plots with average nightlight intensity on the 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 17 to 6 when one compares a market that lies on top of a road to a market that is $2km$ from a road. This change in average DN intensity implies a drop in GDP of 20%, when using the elasticity reported in Henderson et al. (2012). 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 measurement 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 3161 markets (compared to 13,314 using a minimum of one MODIS pixels at 1 km buffer). We then compare the elasticity of market area and average DN value to distance from the closest road in Figure 14 (where panels 1 and 2 replicate DN 33 Markets and MODIS markets from the previous two figures). The MODIS markets that impose a $1km$ minimum area (panel 3) 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.

4.2 Intra-urban versus Inter-urban Market Access

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. As an example of this utility, Figure 15 shows a map of MODIS landcover-based markets, at different

buffers, for New Dehli. The gray polygon represents the $8km$ buffered market that contains the centroid of the city, which we refer to as a super-market. This $8km$ buffer contains many smaller sub-markets buffered at $4km$, $2km$, or $1km$. New Delhi has 125 $(4km)$, 340 $(2km)$, and 463 $(1km)$ sub-markets that are contained within its $8km$ buffer. These sub-markets represent smaller urban zones or neighborhoods within the larger metropolitan area. Turning to the country as a whole, Figure 16 reports the average number of $i = \{1, 2, 4\}km$ sub-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 sub-markets, the presence of these markets is a general phenomenon detectable via landcover-based market definitions. For example, an average of 1.8 $1km$ buffered MODIS markets lie within super-markets defined at a $4km$ buffer, and an average of 3.8 $1km$ buffered markets lie within $8km$ super-markets.

How might we deploy data on landcover-based sub-markets and super-markets? One application would be to detect the consequences of infrastructure developments on intra-urban areas within larger buffered markets. In the spirit of such analysis, we examine the average distances to other sub-markets within given super-markets, which are reported in Figure 17. Consider MODIS markets. Within a $8km$ buffer, the average distance between $1km$ sub-markets is 39.5 kilometers, indicating that the typical $8km$ buffered super-market is an economic region unto itself that 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 4.8 kilometers, which indicates that at a $4km$ buffer we are dealing with entities that more closely represent collections of interconnected neighborhoods. The contrast in sub-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.

As a final exercise to illustrate the value of distinguishing urban sub-markets within their respective super-markets, we calculate market access (e.g., Donaldson and Hornbeck 2016) for our landcover-based definitions. 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). Following Donaldson and Hornbeck (2016), we exclude the own market in the summation. 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 $k = \{2, 4, 8\}km$. Table 18 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, 2.5%, 6.5% and 21.4% 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 5.5%, 14.0% and 36.1%. 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, we are now in a position to provide a much higher resolution characterization of how reduced travel times and 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. These advances are likely to be particularly valuable for analysis in developing countries, where geographically disaggregated data are available infrequently and inconsistently.

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 urban markets, 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 cities. The creation of new methods for integrating alternative sources of satellite imagery is a promising avenue for research.

We have provided just one of many possible applications of high-resolution satellite imagery: to define urban markets in a manner that allows economists to shrink the geographic unit of analysis for virtually any country in the world from the rough equivalent of a U.S. metropolitan area to the rough equivalent of a U.S. neighborhood. With existing analytical tools, these data will make it possible to evaluate the potentially highly spatially heterogeneous economic impacts of investments in infrastructure. With the continents of Asia and Africa in the midst of a multi-trillion dollar investments, the arrival of such capabilities is well timed.

While satellite imagery greatly expands the supply of data amenable to economic analysis, their interpretation is, at this stage, partially limited by the supply of conventionally measured economic quantities. Demand will be particularly high for methods to validate satellite-based measures of economic activity using additional sources of micro data, such as geocoded mobile phone data, economic censuses, and crowdsourced information. 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.

AHLFELDT, G. M., S. J. REDDING, D. M. STURM, AND N. WOLF (2015): "The Economics of Density: Evidence From the Berlin Wall," *Econometrica*, 83, 2127–2189.

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.

CHAUVIN, J. P., E. GLAESER, Y. MA, AND K. TOBIO (2016): "What is Different About Urbanization in Rich and Poor Countries? Cities in Brazil, China, India and the United States," Working Paper 22002, National Bureau of Economic Research.

DAVIS, D. R., J. I. DINGEL, AND A. MISCIO (2018): "Cities, Skills, and Sectors in Developing Economies," *mimeo Columbia University*.

DONALDSON, D. (2018): "Railroads of the Raj: Estimating the Impact of Transportation Infrastructure," *American Economic Review*, 108, 899–934.

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. (2007): "Urban Evolutions: The Fast, the Slow, and the Still," *American Economic Review*, 97, 197–221.

——— (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.

——— (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.

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

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.

GABAIX, X. (1999): "Zipf's Law for Cities: An Explanation," *The Quarterly Journal of Economics*, 114, 739–767.

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.

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.

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

HARRIS, C. D. (1954): "The, Market as a Factor in the Localization of Industry in the United States," *Annals of the Association of American Geographers*, 44, 315–348.

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, J. V. AND A. J. VENABLES (2009): "The dynamics of city formation," *Review of Economic Dynamics*, 12, 233 – 254.

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.

KRUGMAN, P. (1991): "Increasing Returns and Economic Geography," *Journal of Political Economy*, 99, 483–99.

LUCAS, R. E. AND E. ROSSI-HANSBERG (2003): "On the Internal Structure of Cities," *Econometrica*, 70, 1445–1476.

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

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).

REDDING, S. AND A. VENABLES (2004): "Economic geography and international inequality," *Journal of International Economics*, 62, 53–82.

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.

ROSSI-HANSBERG, E. AND M. L. J. WRIGHT (2007): "Establishment Size Dynamics in the Aggregate Economy," *American Economic Review*, 97, 1639–1666.

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.

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.

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.

WULDER, M. A., J. C. WHITE, T. R. LOVELAND, C. E. WOODCOCK, A. S. BELWARD, W. B. COHEN, E. A. FOSNIGHT, J. SHAW, J. G. MASEK, AND D. P. ROY (2016): "The global Landsat archive: Status, consolidation, and direction," *Remote Sensing of Environment*, 185, 271 – 283, landsat 8 Science Results.

Appendix A

This 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 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 built up activity and appear unlit only because of blooming. We then proceed with the classification.

The steps of the methodology are as follows:

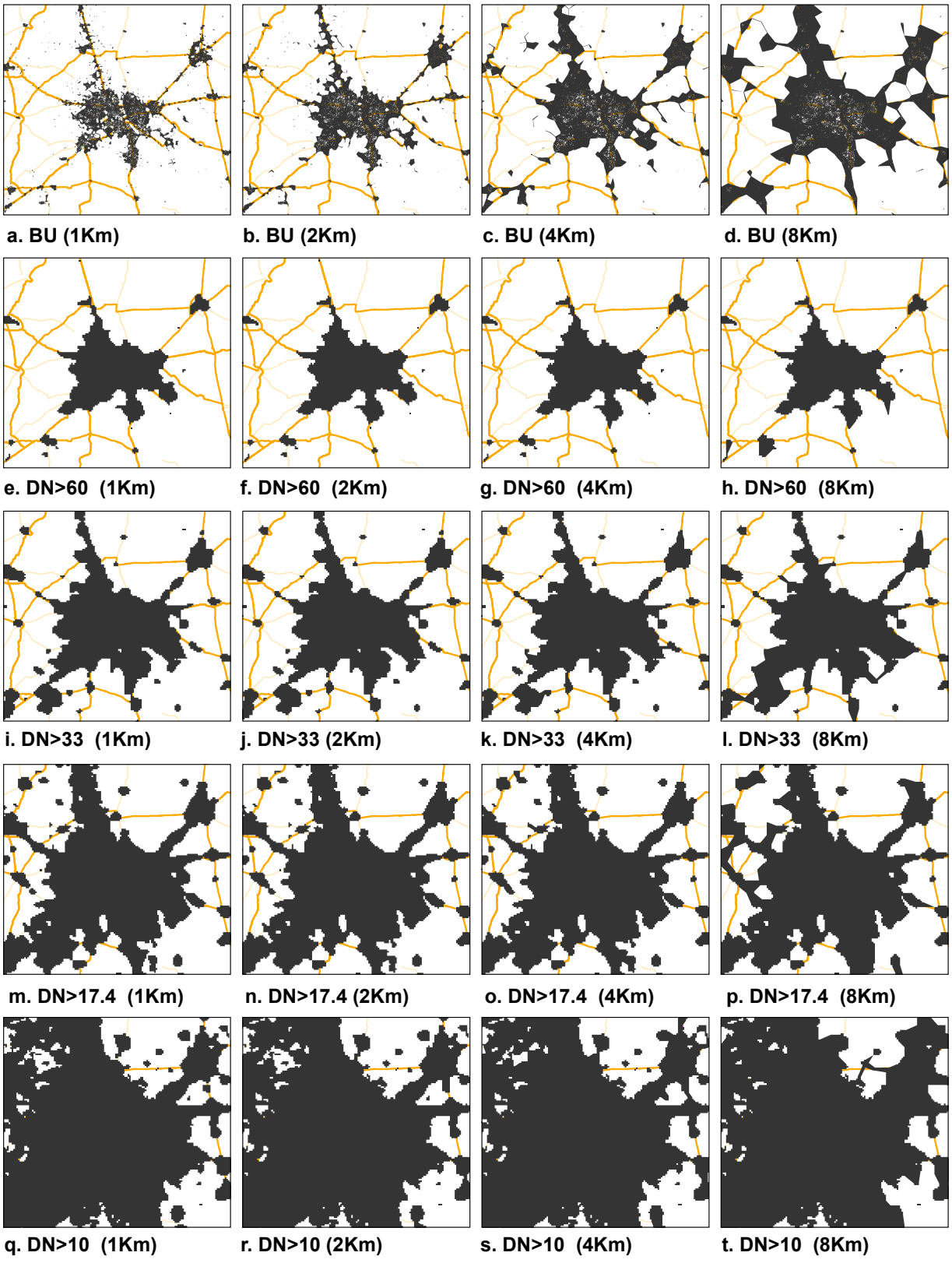
1. Designate a pixel as *builtup* if its DN exceeds a threshold. This threshold is set at the 95^{th} percentile of pixels in the training set, which is 17.4 across all India but ranges is allowed to vary across hex-cells (discussed below).
2. Re-classify a builtup pixel as *not builtup* if the Landsat index bands (NDVI, NDWI, NDBI) indicate presence of water, dense vegetation or not built-up 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). The Otsu algorithm 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%.[1] Note that this accuracy rate exceeds the MODIS classification accuracy by 2.5% in India; see Table 6 of Goldblatt et al. (2018).

[1] Accuracy rate is defined as the sum of true positives and true negatives divided by the total sample.

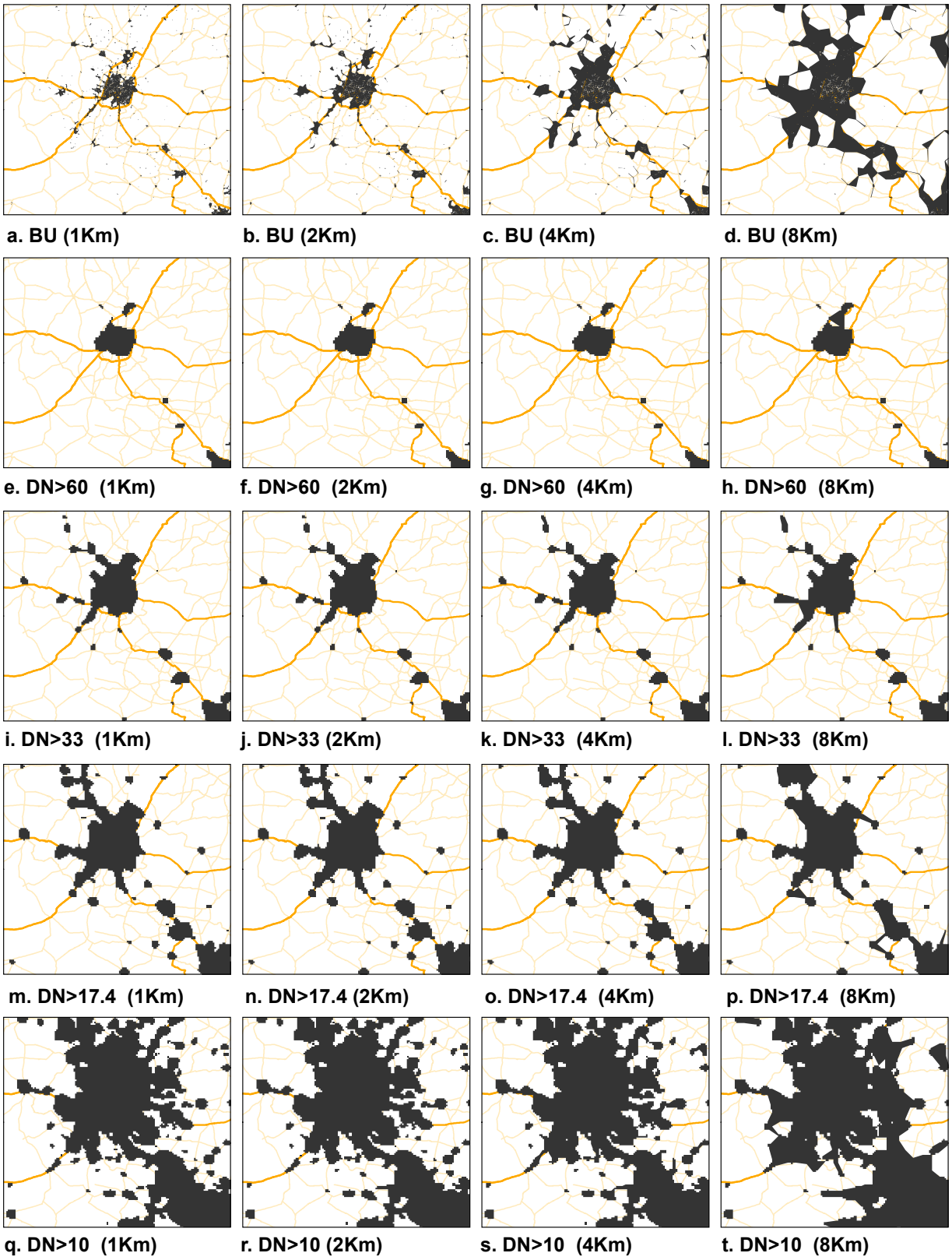
Figures

Figure 1: Delhi, Alternative Market Definitions



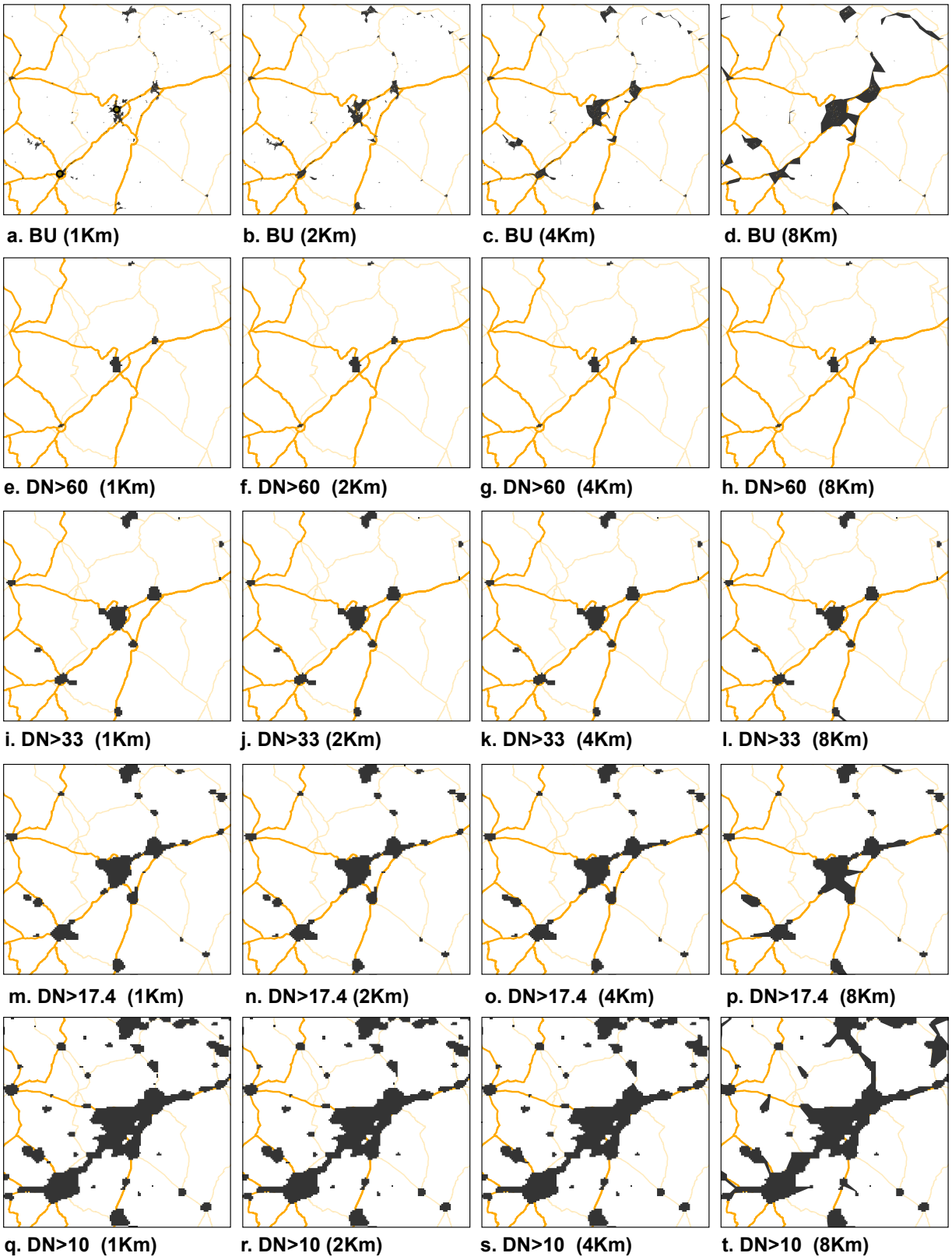
The figure displays markets around New Delhi. Row 1 displays landcover-based markets using the MIX layer. Row 2-5 displays nightlight-based markets.

Figure 2: Ahmedabad, Alternative Market Definitions



The figure displays markets around Ahmedabad. Row 1 displays landcover-based markets using the MIX layer. Row 2-5 displays nightlight-based markets.

Figure 3: Ajmer, Alternative Market Definitions



The figure displays markets around Ajmer. Row 1 displays landcover-based markets using the MIX layer. Row 2-5 displays nightlight-based markets.

Figure 4: Number of Urban Markets, by Market Definition

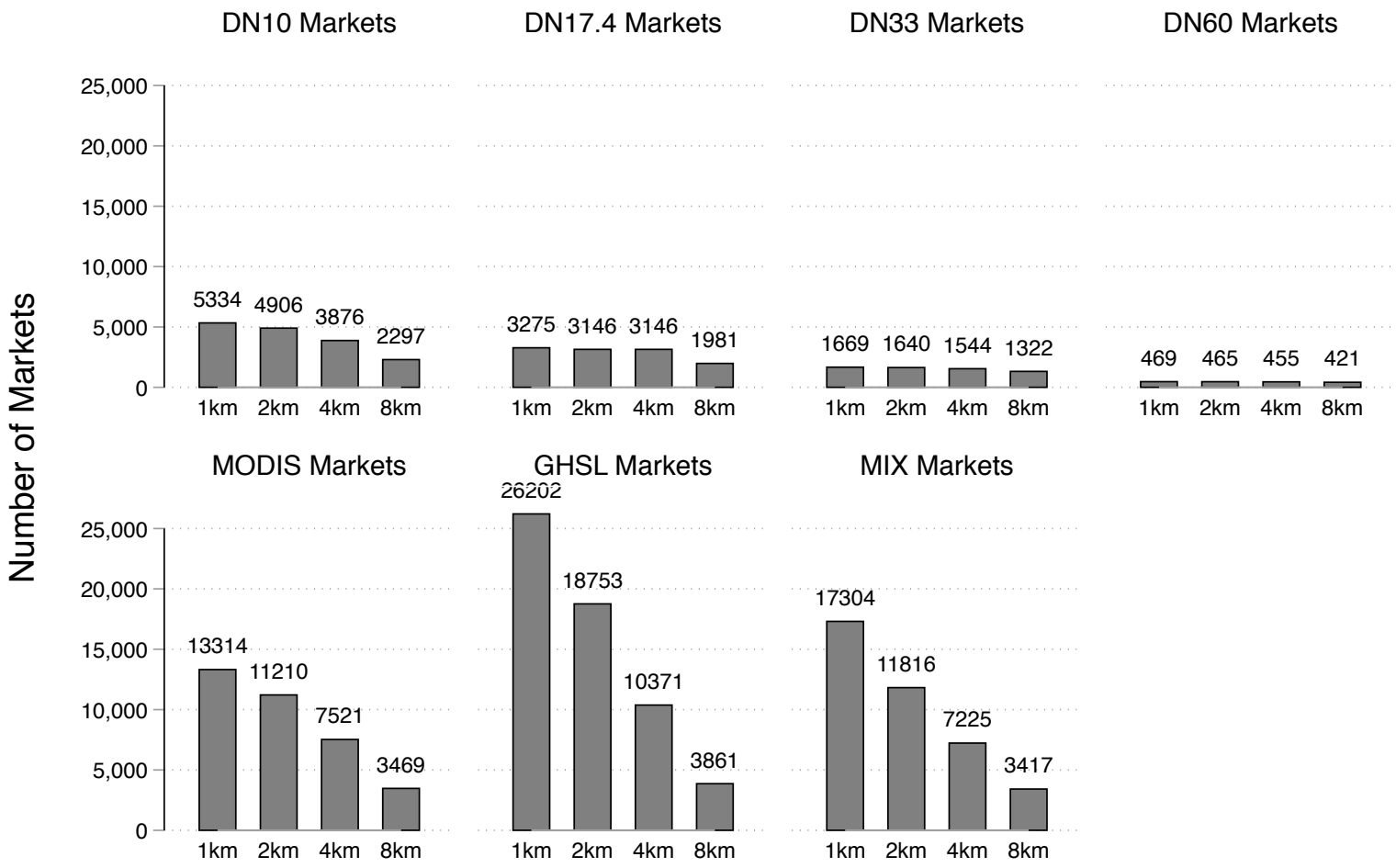


Figure reports the number of markets by market definition

Figure 5: Average Land Area, by Market Definition

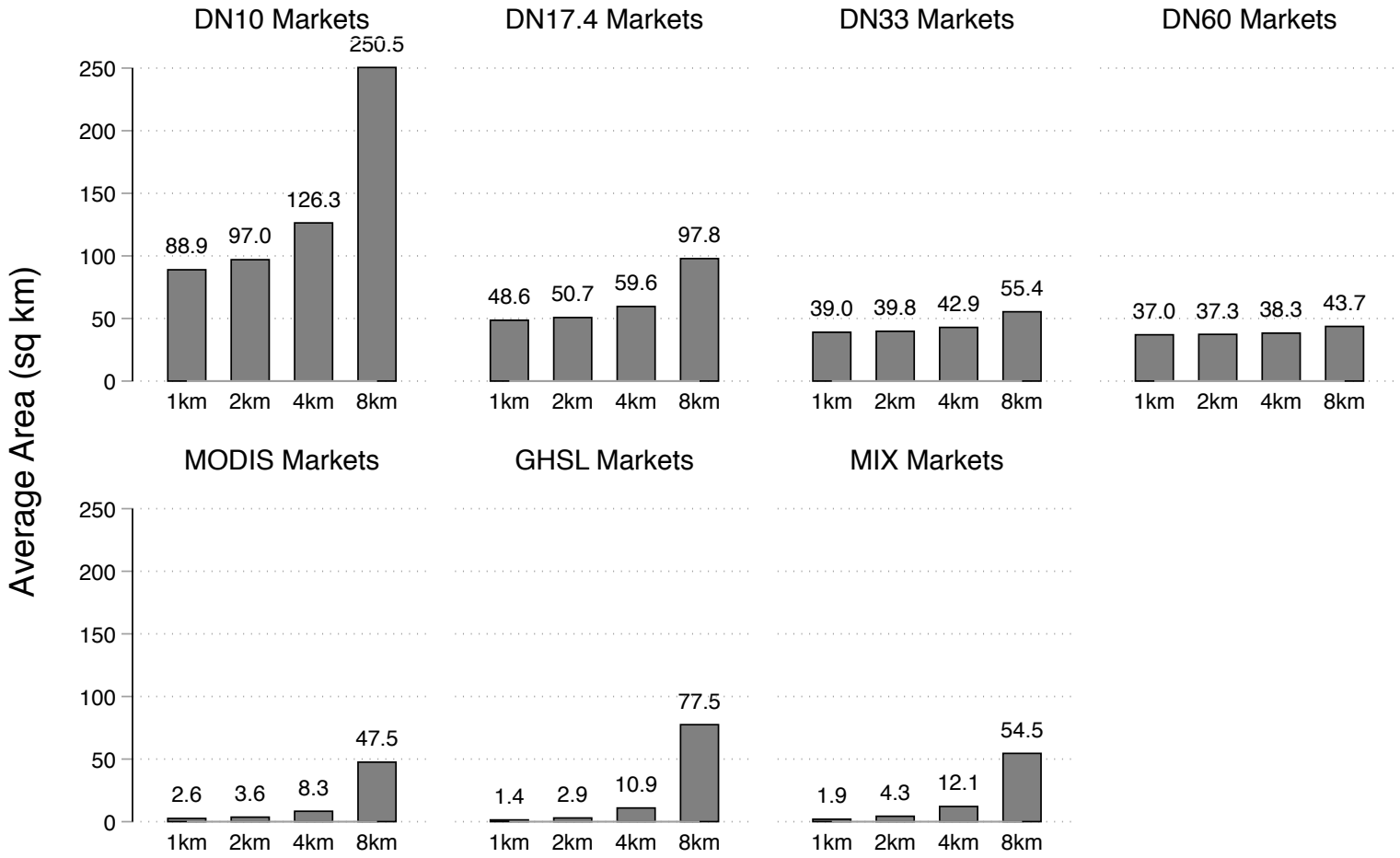


Figure reports the average area by market definition

Figure 6: Distribution of Nightlight DN Values, by Market Definition

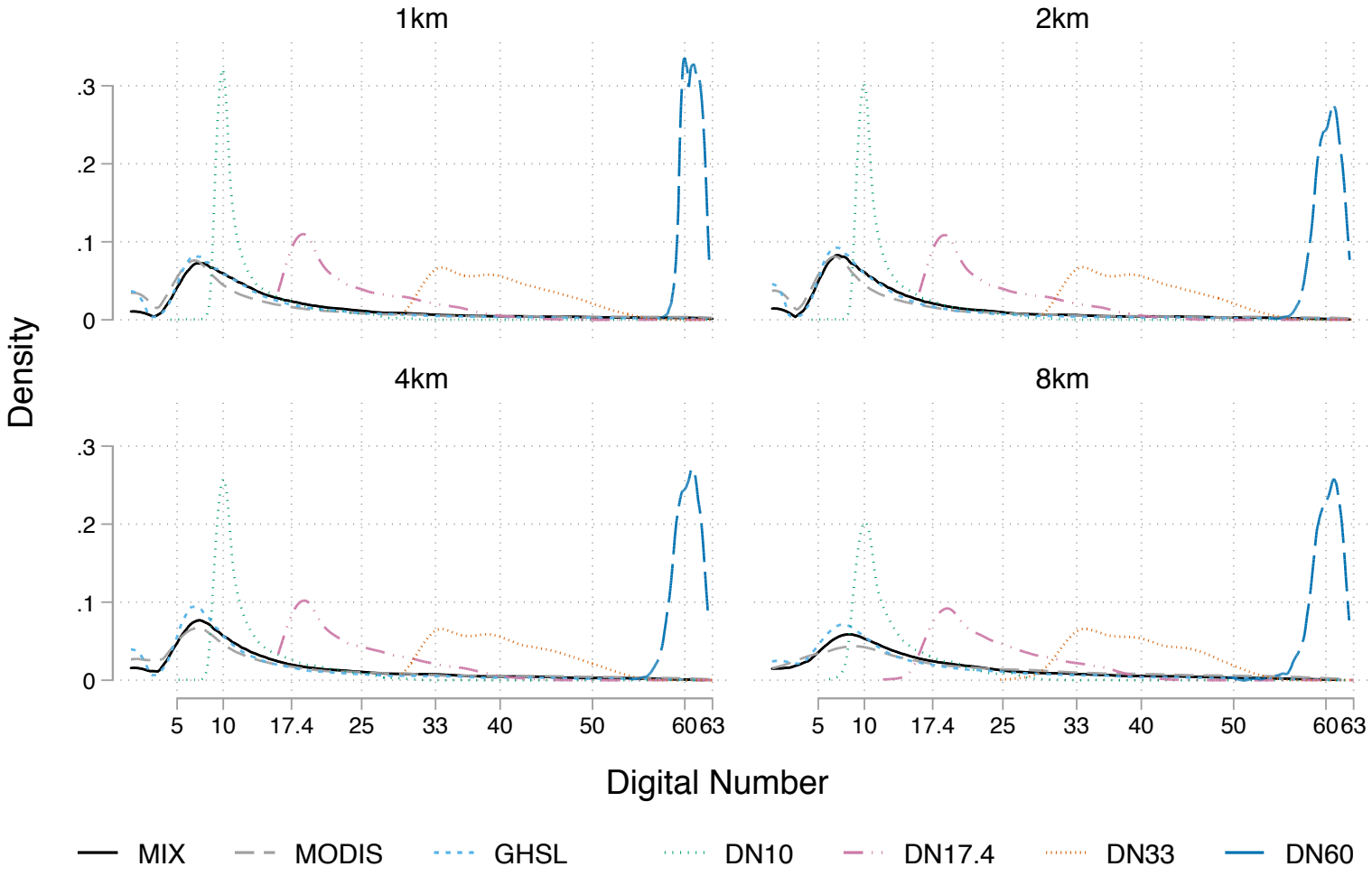
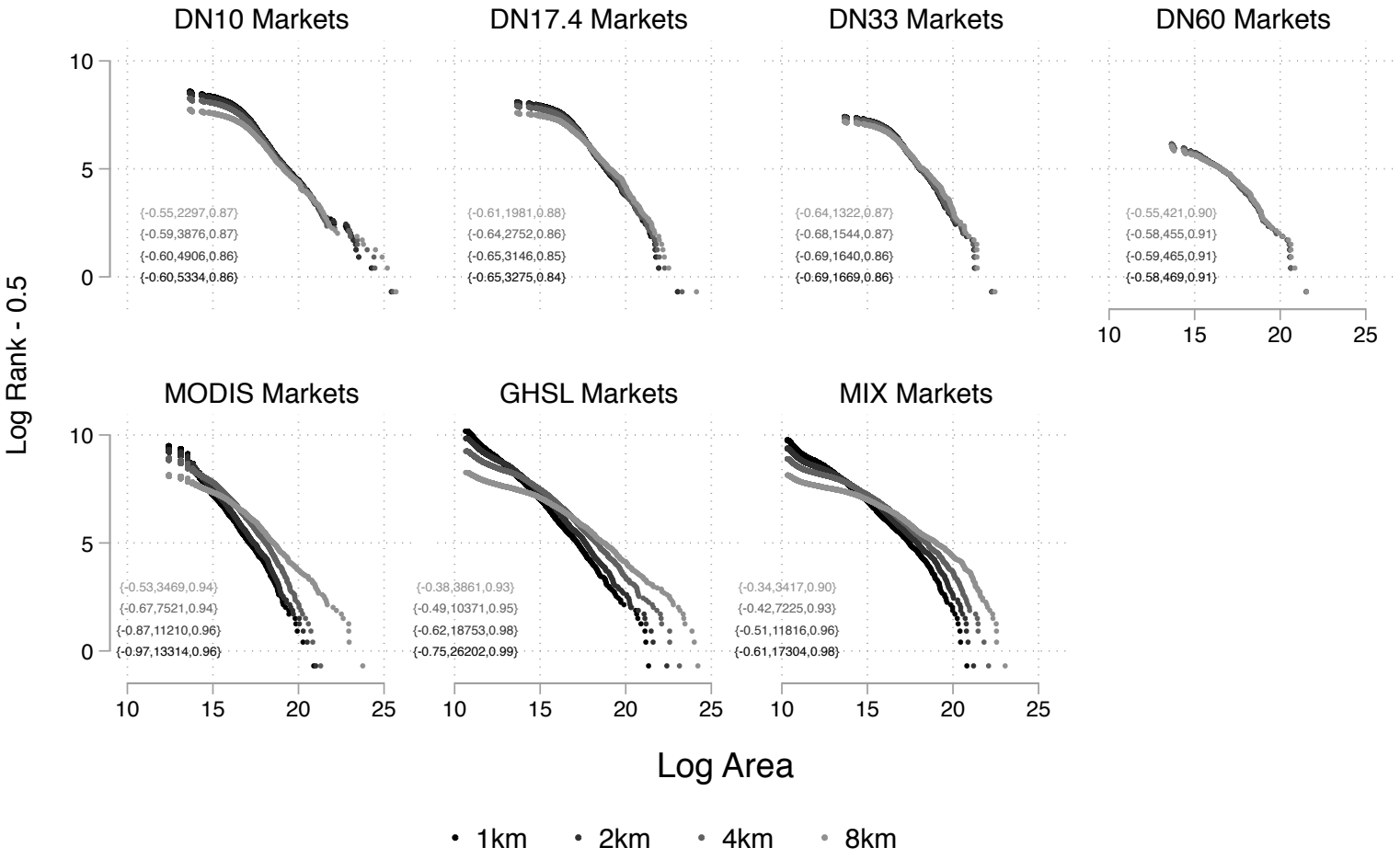
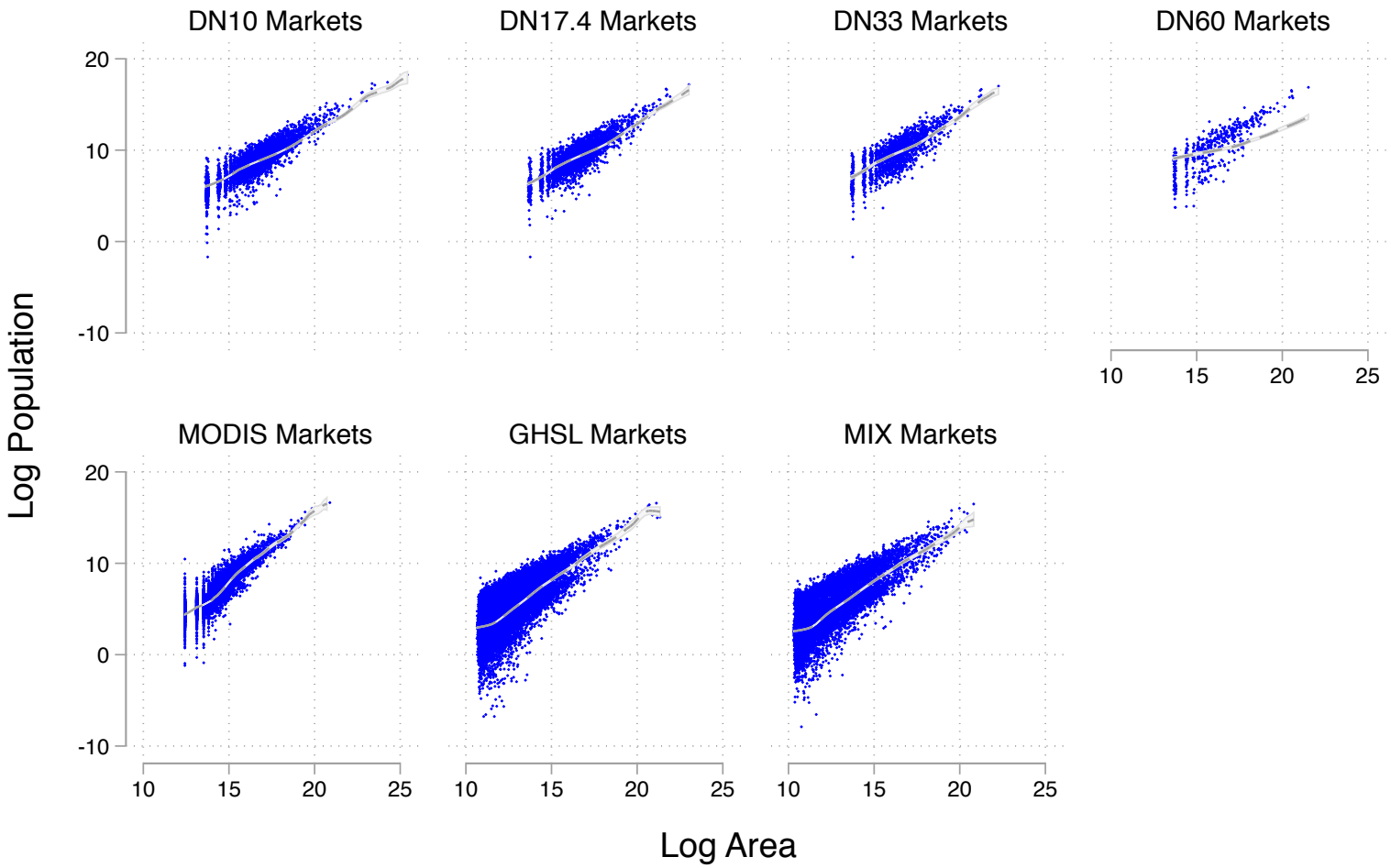


Figure 7: Land Area-Rank Relationship, by Market Definition



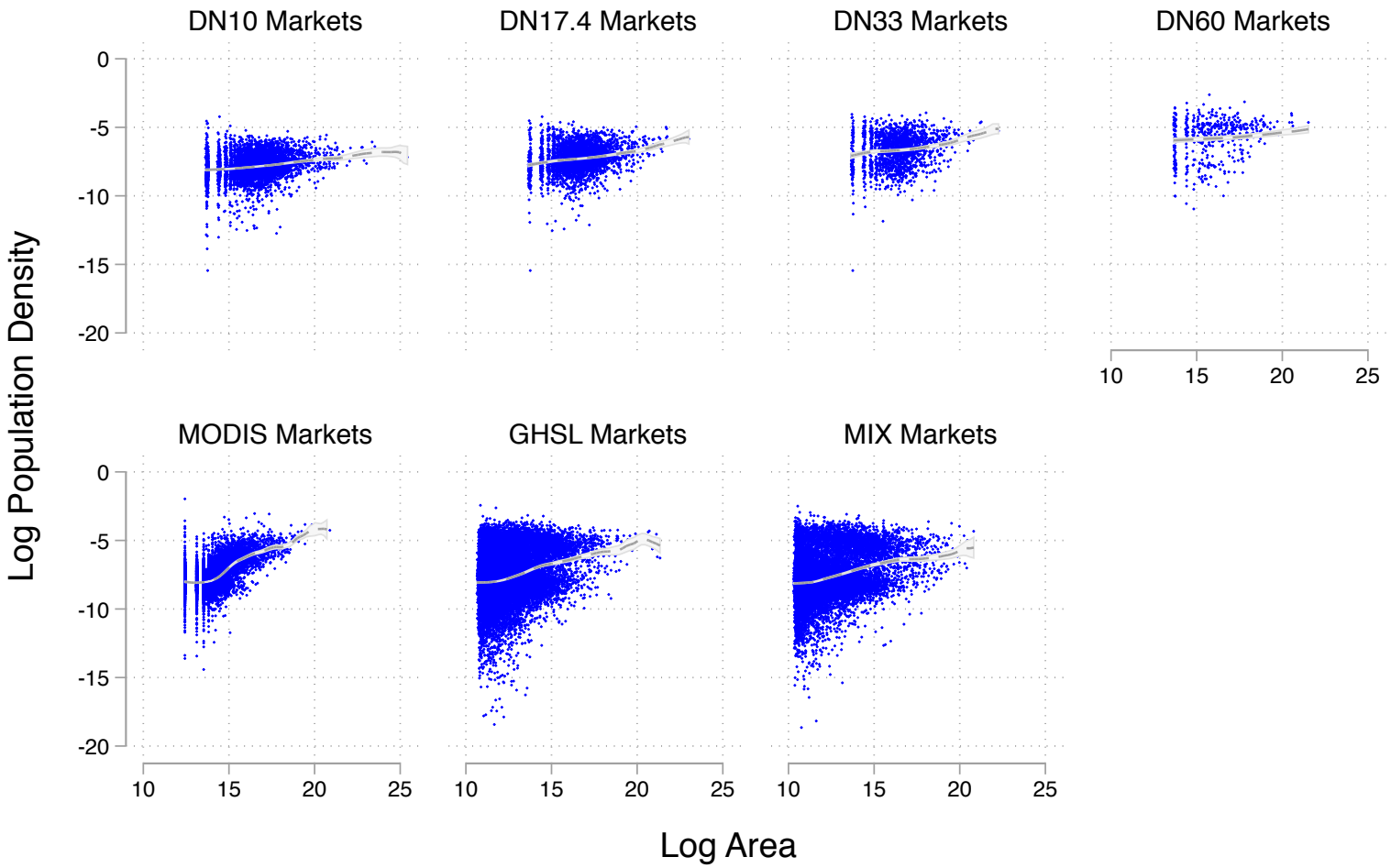
The {slope, number of observations, r-squared} reported for the regressions: log(rank - 0.5) = constant + slope*log(area) + error

Figure 8: Population versus Land Area, by Market Definition (1km Buffer)



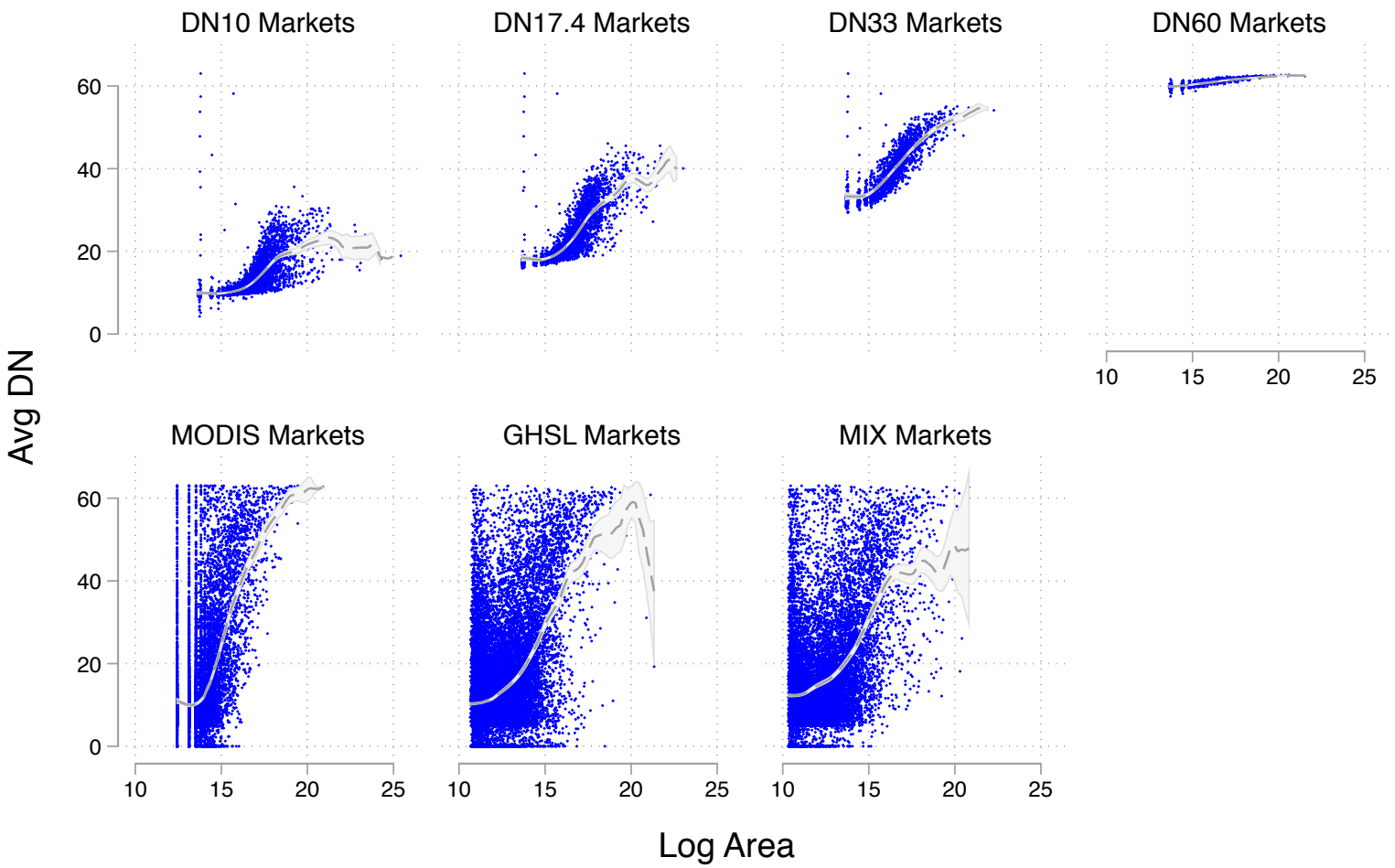
Markets in this figure are buffered at 1km. Population data from WorldPop.

Figure 9: Population Density versus Land Area, by Market Definition (1km Buffer)



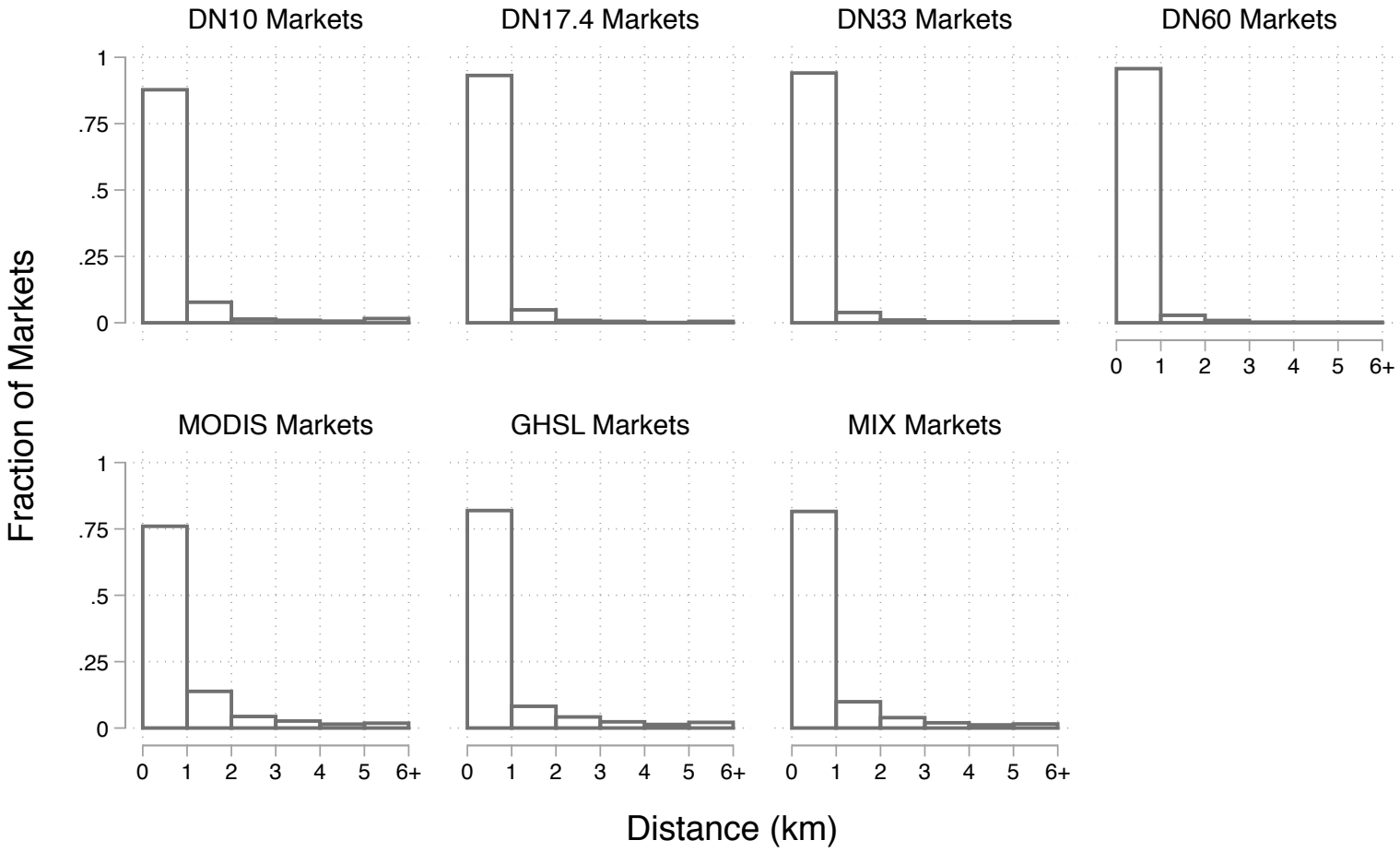
Markets in this figure are buffered at 1km. Population data from WorldPop.

Figure 10: Average DN Intensity versus Land Area, by Market Definition (1km Buffer)



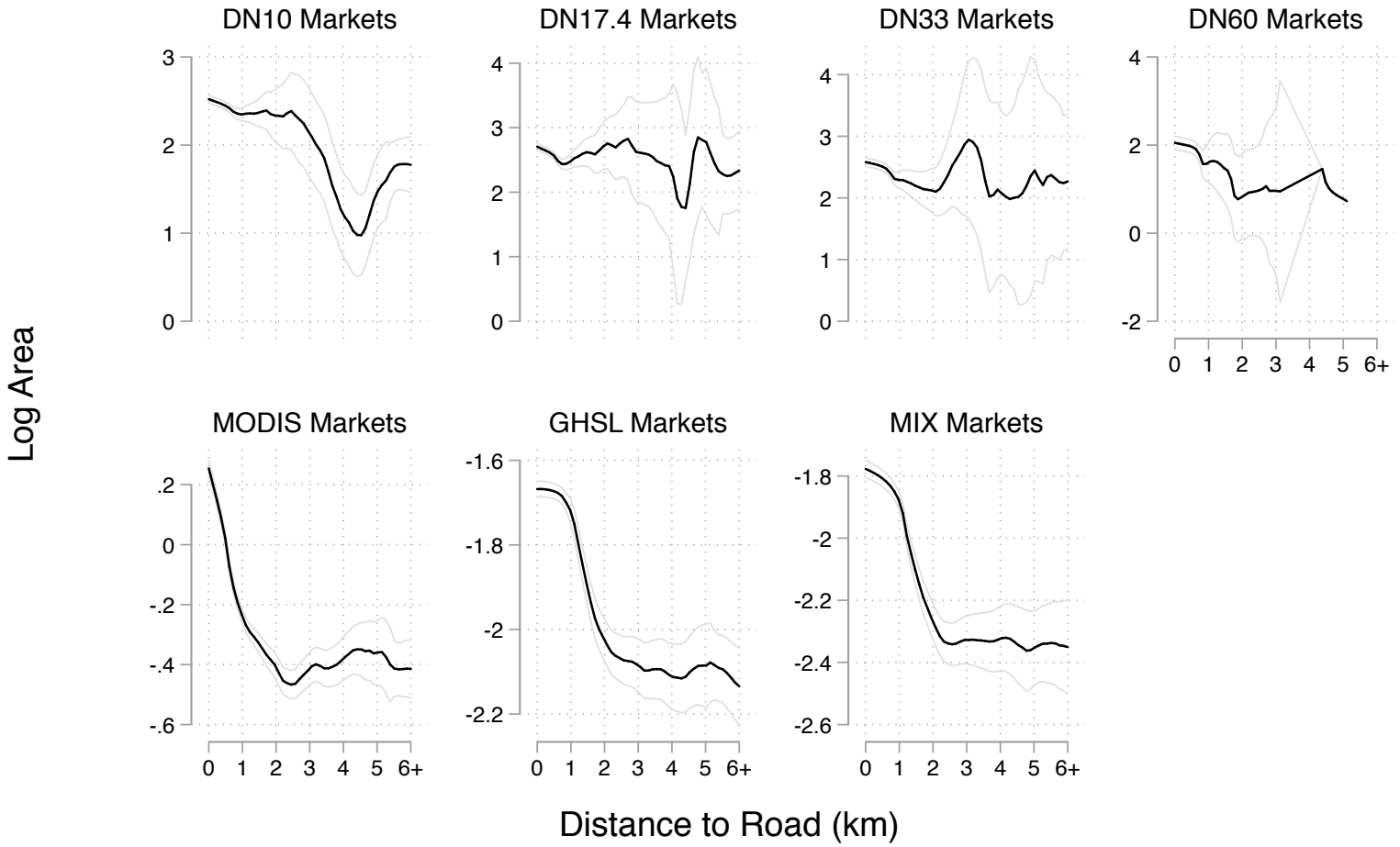
Markets in this figure are buffered at 1km

Figure 11: Distance to Nearest Road, by Market Definition (1km Buffer)



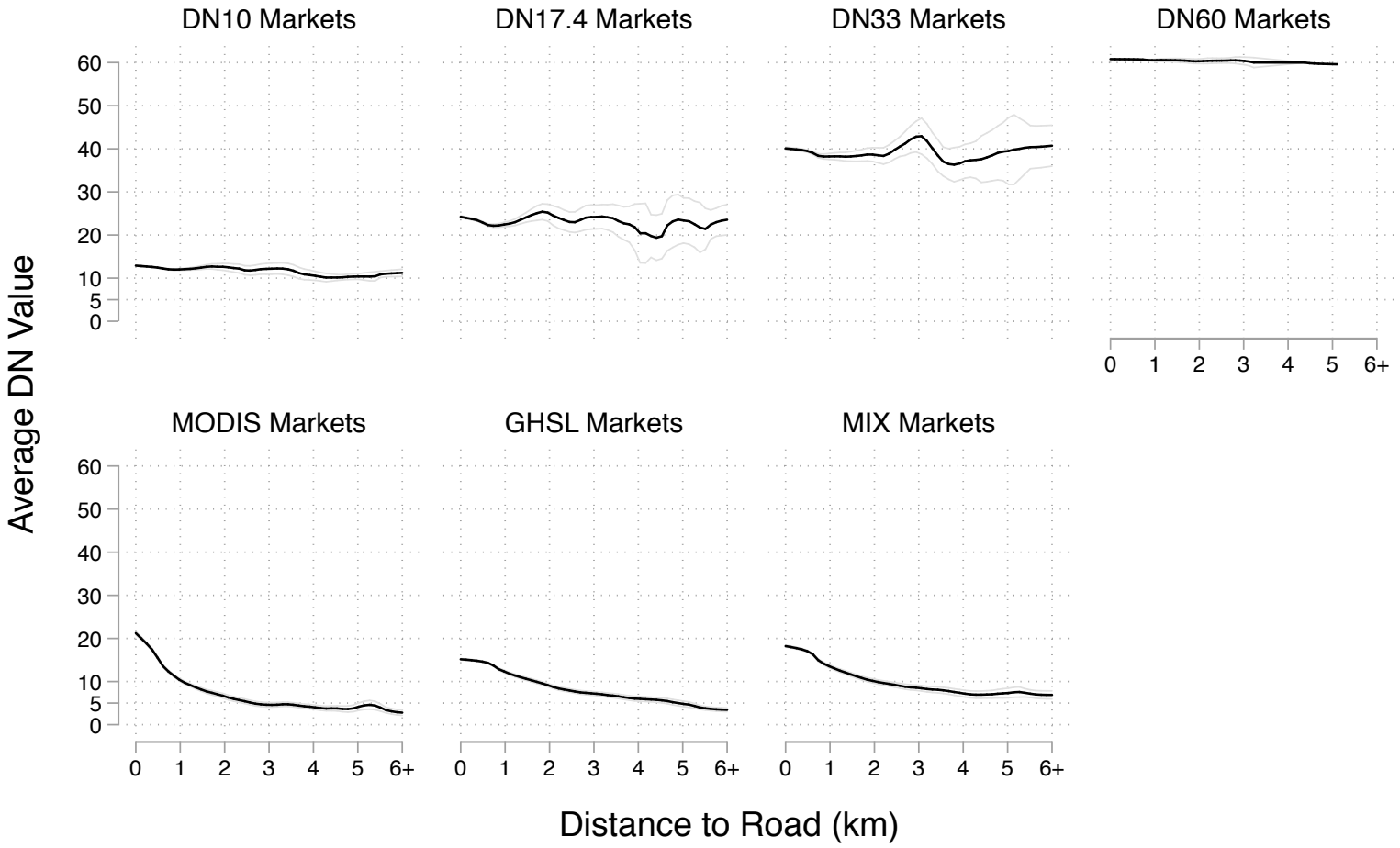
Shortest distance from market centroid to a primary, secondary or tertiary road.
Road data from India's current network on OpenStreetMaps. Markets buffered at 1km.

Figure 12: Land Area versus Distance to Nearest Road, by Market Definition (1km Buffer)



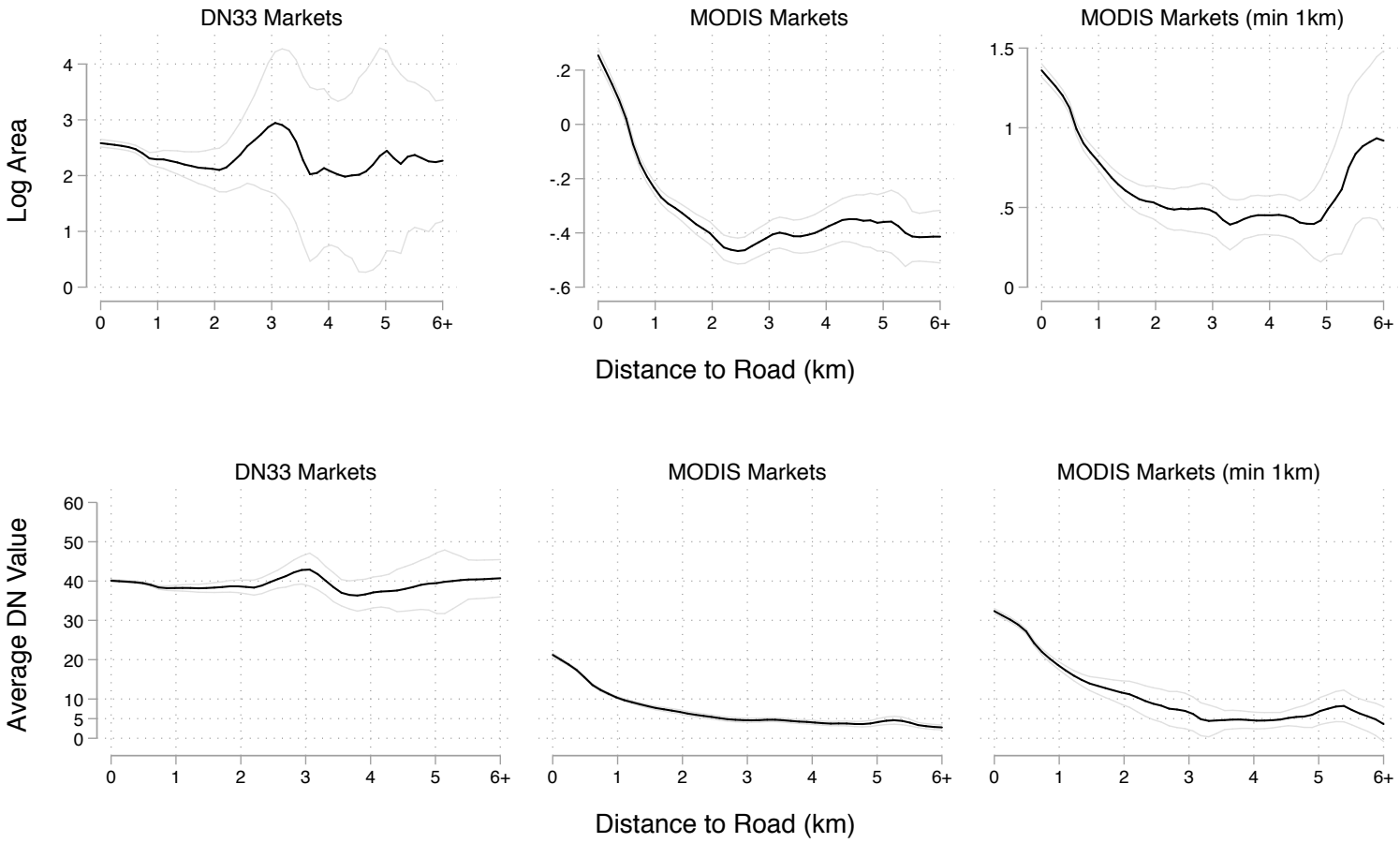
Shortest distance from market centroid to a primary, secondary or tertiary road.
Road data from India's current network on OpenStreetMaps. Markets buffered at 1km.

Figure 13: Average DN versus Distance to Nearest Road, by Market Definition (1km Buffer)



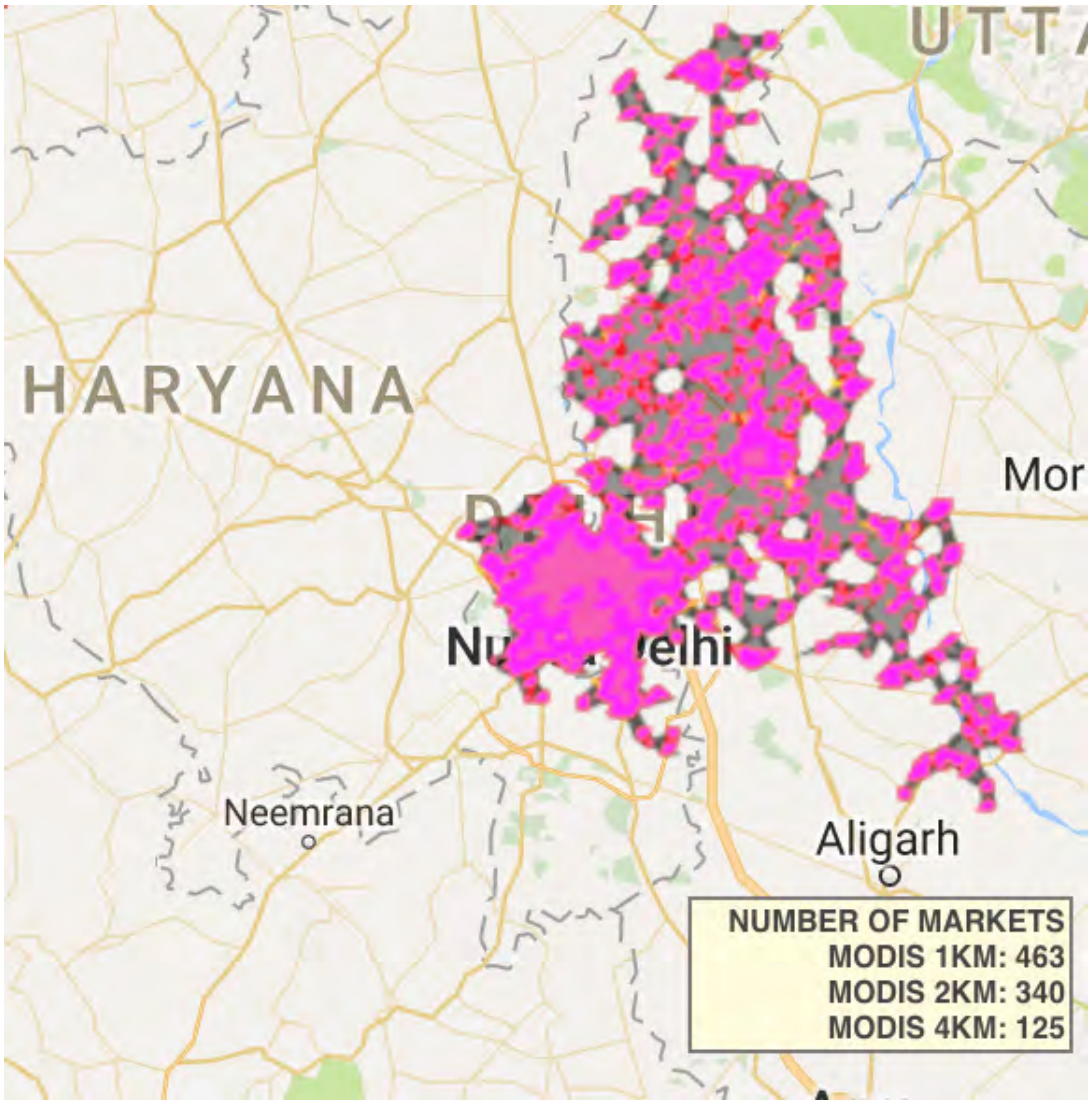
Shortest distance from market centroid to a primary, secondary or tertiary road.
Road data from India's current network on OpenStreetMaps. Markets buffered at 1km.

Figure 14: DN33 Markets versus Coarser MODIS Markets (1km minimum area)



Shortest distance from market centroid to a primary, secondary or tertiary road.
Road data from India's current network on OpenStreetMaps. Markets buffered at 1km. First two panels repeated from previous figure.

Figure 15: MODIS Landcover-Based Markets within New Delhi $8km$ Buffer



The figure displays $\{1, 2, 4\}km$ markets that lie with the New Delhi $8km$ buffer. The grey polygon is the $8km$ buffer. The pink polygons are the $1km$ markets.

Figure 16: Sub-markets within Super-markets, by Landcover-Based Market Definition

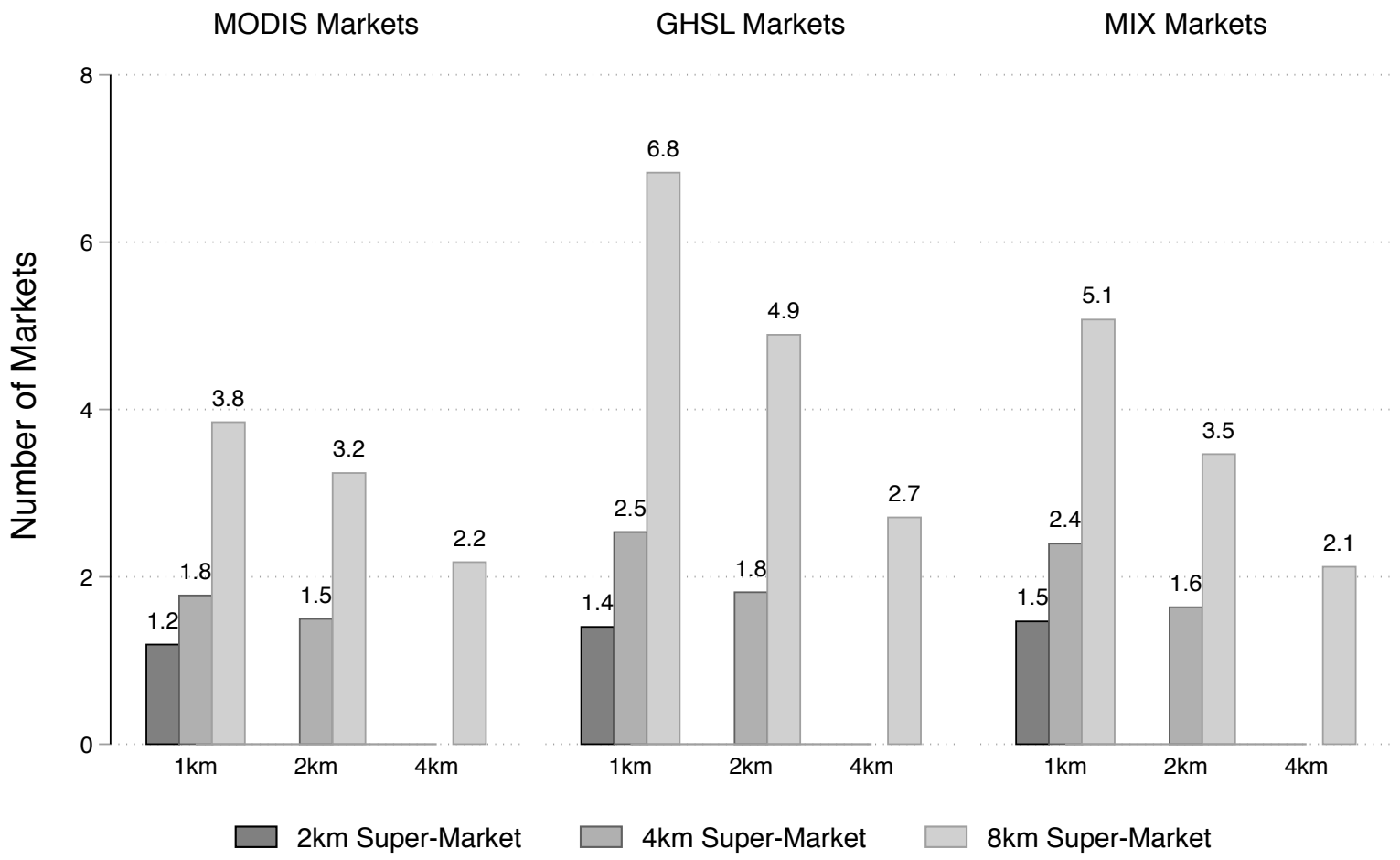


Figure reports the average number of 'sub-markets' that lie with a 'super-market'.

Figure 17: Distance within Super-Markets, by Landcover-Based Market Definition

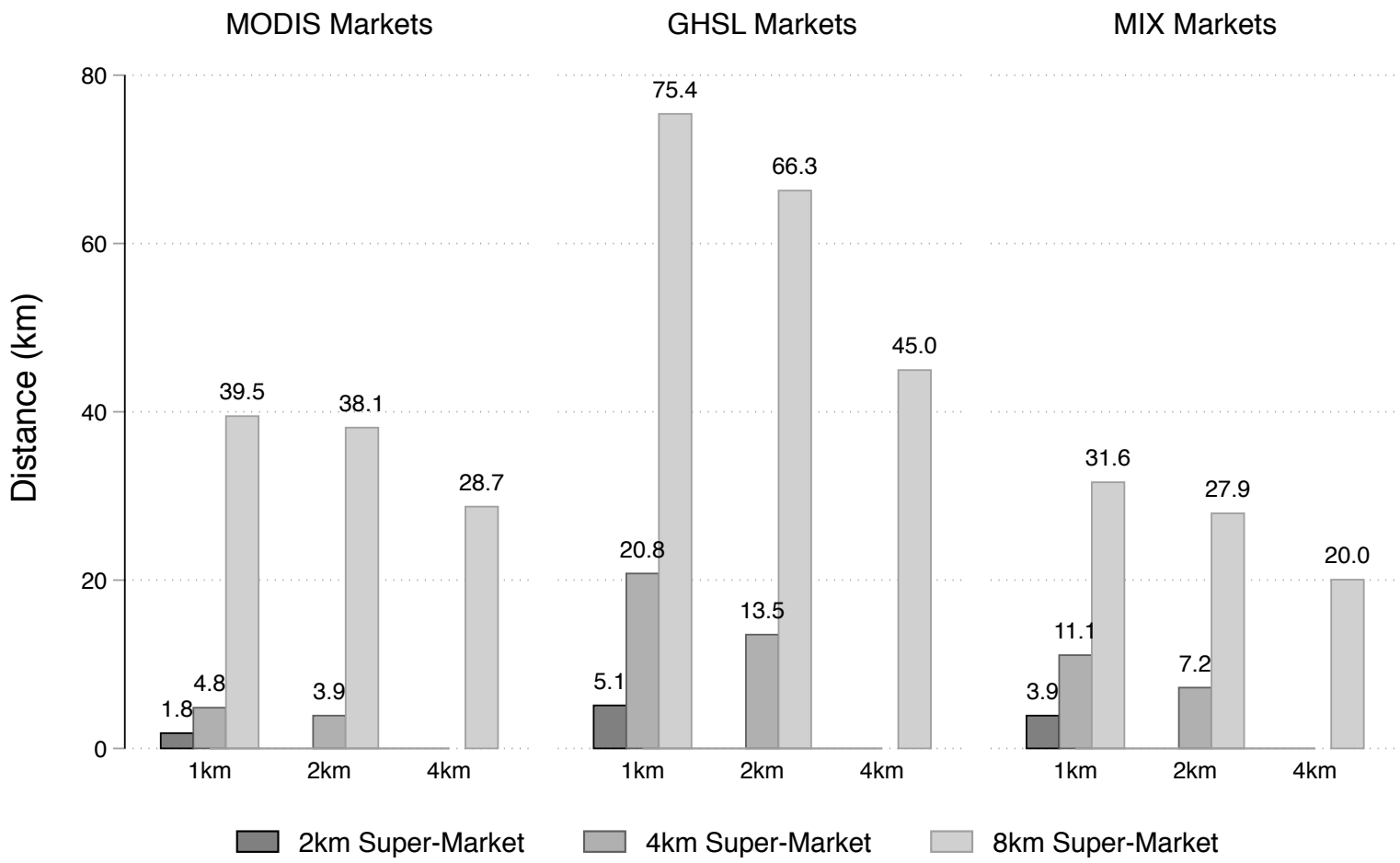


Figure reports the average distance to other 'sub-markets' within the 'super-market'.

Figure 18: Share of Market Access Accounted for by Other Sub-markets within Super-Market

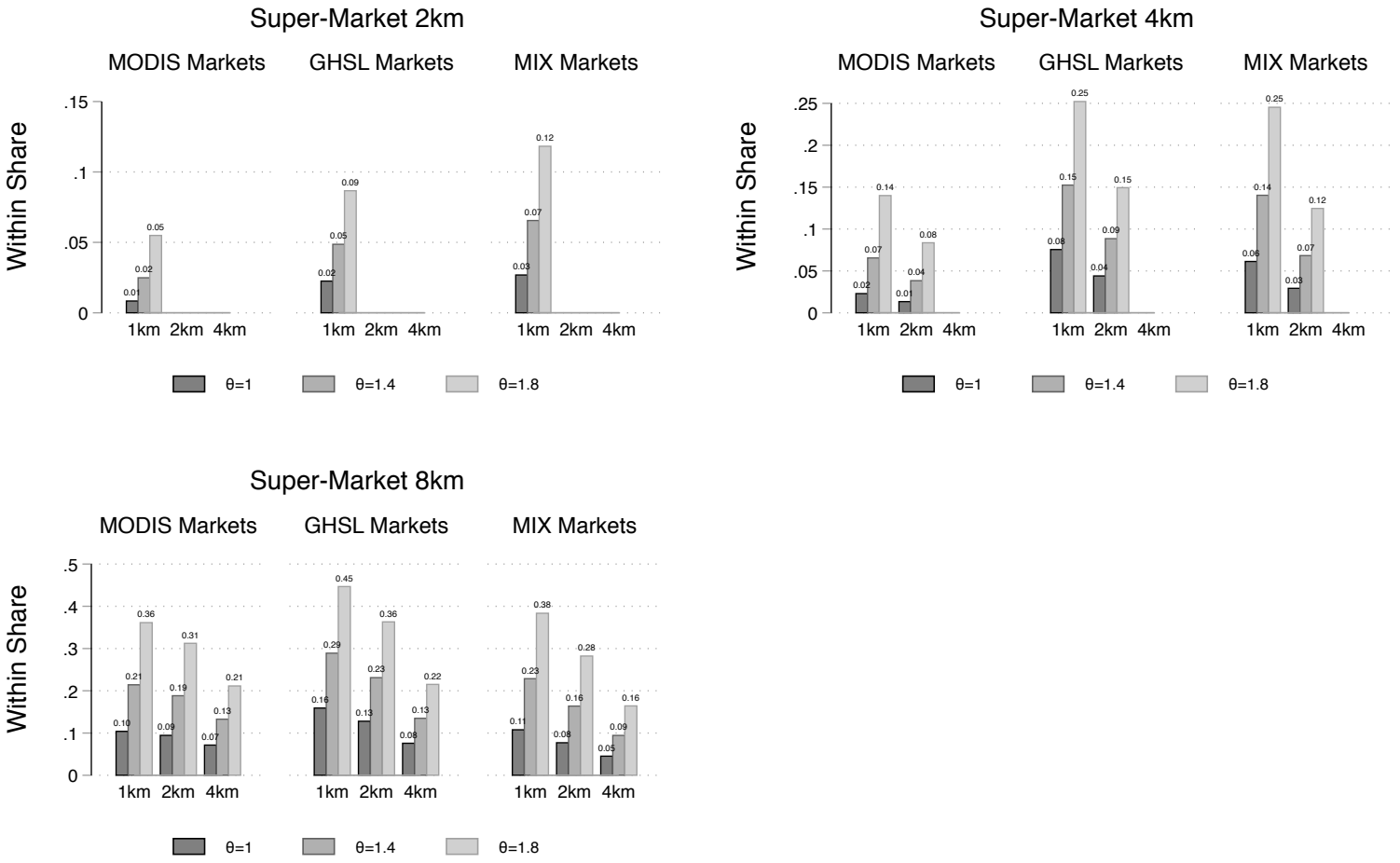
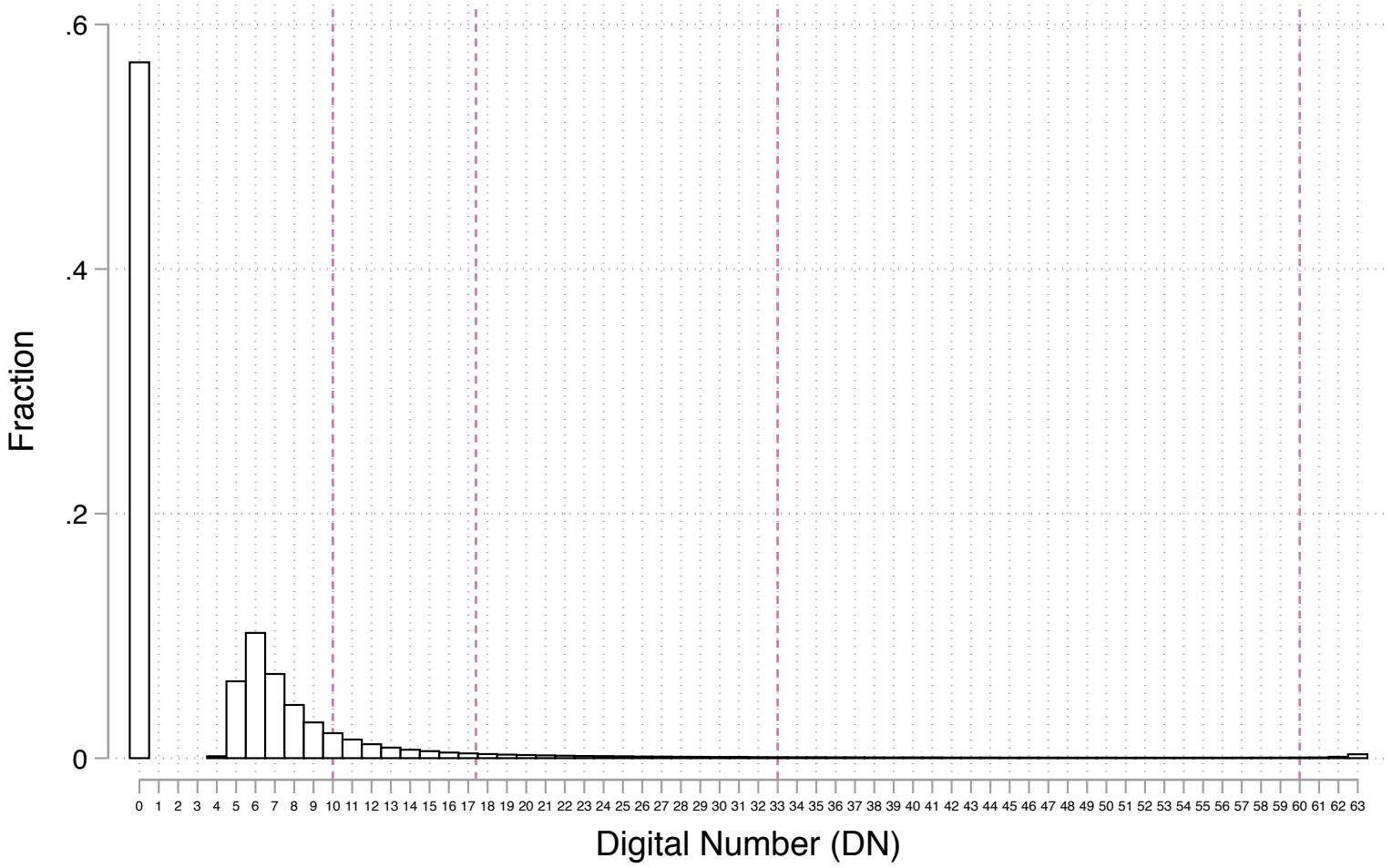


Figure denotes the average share of market access accounted by 'sub-markets' within a 'super-market', for different values of θ.

Online Appendix Figures and Tables

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



Vertical line reflects the 90th, 95th, 99th and 99.5th percentiles of the DNs. Histogram is a 3% random sample of India's total pixels.

Figure A2: Combining Polygons to Form Markets

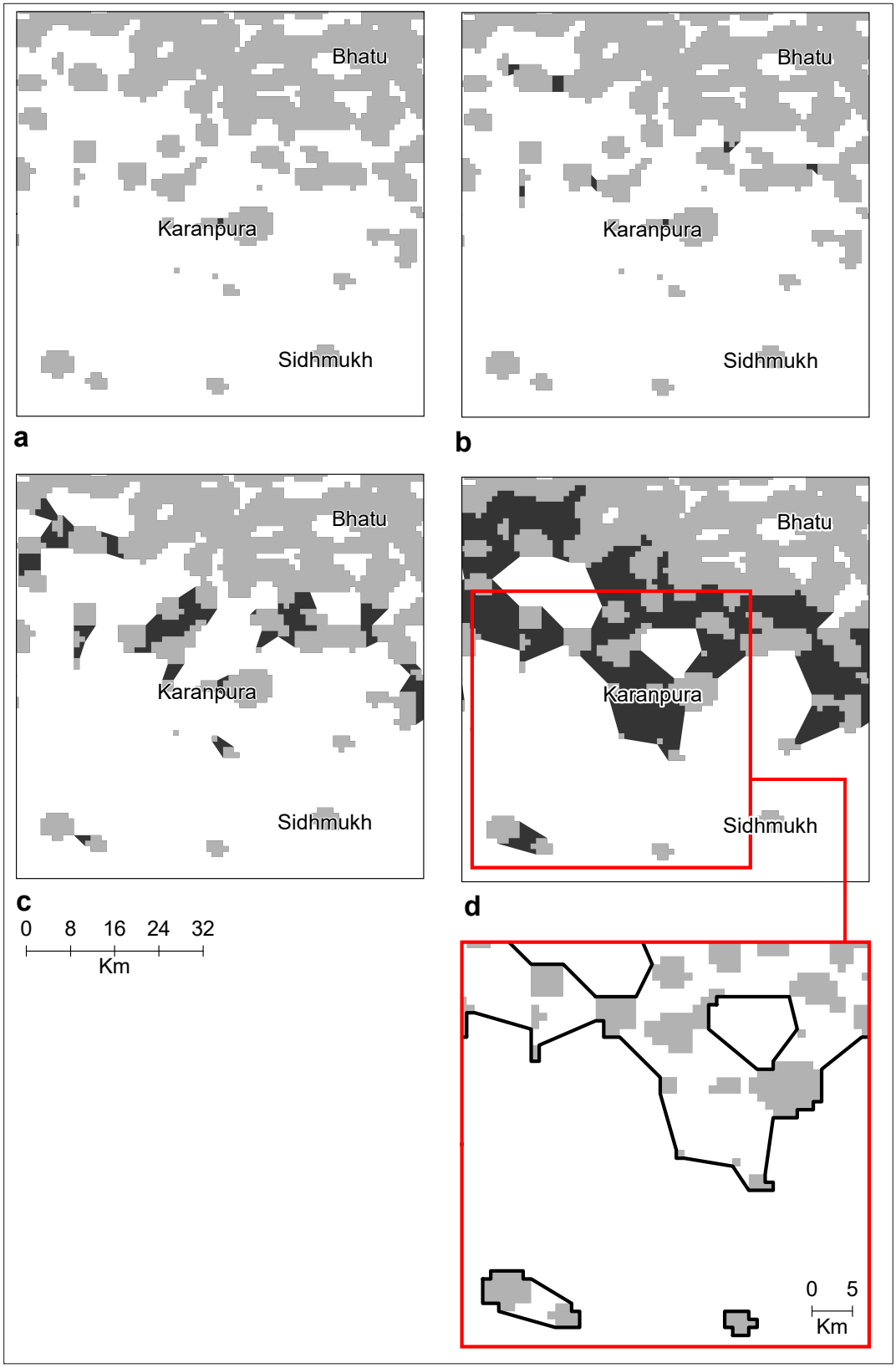


Figure A3: Distribution of Land Area, by Market Definition

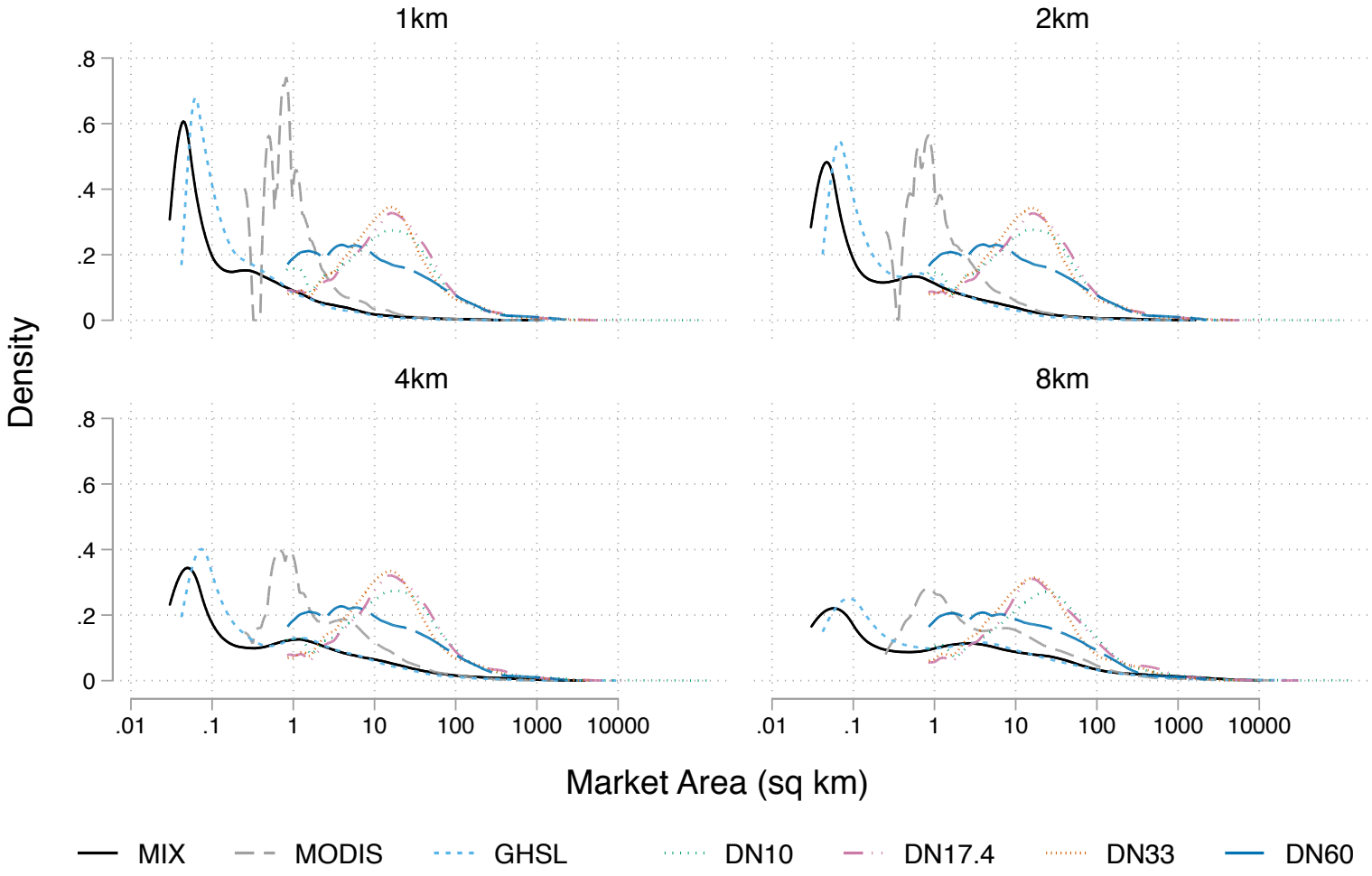


Table A1: Administrative Areas in India, 2011 Census

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

Source: Census 2011