Online Appendix

Cities in Bad Shape: Urban Geometry in India

Mariaflavia Harari

This Appendix is organized as follows. Section A discusses in detail the data employed in the paper. Section B presents a formal derivation of the spatial equilibrium model outlined in Section 4 in the paper. Section C outlines an alternative model of within-city spatial equilibrium to shed light on the rent gradient in a city with irregular shape. Section D provides an analytical derivation of the bias in the estimation of the OLS impacts of city shape. Section E discusses an alternative version of the shape instrument, used in one of the robustness checks, that does not rely on projecting historical population. Section F describes the procedure used to detect employment sub-centers to obtain the estimates in Table 10.

A. Data

Below I provide details on the sources and methods employed to assemble my dataset. I start by describing how I retrieve urban footprints (Section A1) and measure their geometric properties (Section A2). Next, I discuss population, wages, and rents (Section A3). In Section A4 I discuss the control variables employed in Table A2 and in the robustness checks in Tables 6, A10, A11, and A12. In Section A5 I discuss variables related to infrastructure and transit. In Section A6 I discuss the remaining data sources - geographic constraints used for the construction of the instrument and variables employed for heterogeneous effects analyses, including firms' addresses, slum population, and floor area ratios.

A1. Urban footprints

The first step in constructing the dataset is to trace the footprints of Indian cities at different points in time and measure their geometric properties. The boundaries of urban footprints are retrieved from two sources. The first is the U.S. Army India and Pakistan Topographic Maps, a series of detailed maps covering the entire Indian subcontinent at a 1:250,000 scale (U.S. Army Map Service 1955-). These maps consist of individual topographic sheets. I geo-referenced each of these sheets and manually traced the reported perimeter of urban areas, which are clearly demarcated. These maps are from the mid-50s, but no specific year of publication is provided. For the purposes of constructing the city-year panel, I label these data as 1950 and match them with Census data from 1951.

The second source is the DMSP/OLS Night-time Lights dataset (National Geophysical Data Center 1992-). This consists of night-time imagery recorded by satellites from the U.S. Air Force Defense Meteorological Satellite Program (DMSP) and reports the recorded intensity of Earth-based lights, measured by a six-bit number (ranging from 0 to 63). This data is reported for every year between 1992 and 2010, with a resolution of 30 arc-seconds (approximately 1 square kilometer). The use of the DMSP-OLS dataset for delineating urban areas is quite common in urban remote sensing (Henderson et

al. 2003; Small, Pozzi, and Elvidge 2005; Small and Elvidge 2013). The methodology is the following: first, I overlap the night-time lights imagery with a point shapefile with the coordinates of Indian settlement points, taken from the Global Rural-Urban Mapping Project (GRUMP) Settlement Points dataset (Balk et al. 2006; Center for International Earth Science Information Network et al. 2017). I then set a luminosity threshold (35 in my baseline approach, as explained below) and consider spatially contiguous lighted areas surrounding the city coordinates with luminosity above that threshold. This approach can be replicated for every year covered by the DMSP/OLS dataset.

The choice of luminosity threshold results in a more or less restrictive definition of urban areas, which will appear larger for lower thresholds.[1] To choose luminosity thresholds appropriate for India, I overlap the 2010 night-time lights imagery with available Google Earth imagery. I find that a luminosity threshold of 35 generates the most plausible mapping for those cities covered by both sources.[2] In my main estimation sample (that includes cities covered in 1950 and 2010), the average city footprint occupies an area of approximately 73 square kilometers. In Table A5, I show that the first-stage results and main population IV results are robust to using alternative luminosity thresholds 30 and 40.

Using night-time lights as opposed to alternative satellite-based products, in particular day-time imagery, is motivated by a number of advantages. Unlike products such as aerial photographs or high-resolution imagery, night-time lights cover systematically the entire Indian subcontinent, and not only a selected number of cities. Moreover, they are one of the few sources allowing to detect changes in urban areas over time, due to their yearly temporal frequency. Finally, unlike multi-spectral satellite imagery, night-time lights do not require any sophisticated manual pre-processing and cross-validation using alternative sources.[3]

It is well known that urban maps based on night-time lights tend to inflate urban boundaries, due to "blooming" effects (Small, Pozzi, and Elvidge 2005).[4] This can only partially be limited by setting high luminosity thresholds. In my panel, urban footprints as reported for years 1992-2010 thus reflect a broad definition of urban agglomerations, which typically goes beyond the current administrative boundaries. This contrasts with urban boundaries reported in the U.S. Army maps, which seem to reflect a more restrictive definition of urban areas (although no specific documentation is available). Throughout my analysis, I focus on long differences or include year fixed effects, which amongst other

[1] Determining where to place the boundary between urban and rural areas always entails some degree of arbitrariness, and in the urban remote sensing literature there is no clear consensus on how to set such threshold. It is nevertheless recommended to validate the chosen threshold by comparing the DMSP/OLS-based urban mapping with alternative sources, such as high-resolution day-time imagery, which in the case of India is available only for a small subset of city-years.

[2] For years covered by both sources (1990, 1995, 2000), my maps also appear consistent with those from the GRUMP - Urban Extents Grid dataset, which combines night-time lights with administrative and Census data to produce global urban maps (Balk et al. 2006; Center for International Earth Science Information Network et al. 2017).

[3] An extensive portion of the urban remote sensing literature compares the accuracy of this approach in mapping urban areas with that attainable with alternative satellite-based products, in particular day-time imagery (e.g. Henderson et al. 2003; Small, Pozzi, and Elvidge 2005). This cross-validation exercise has been carried out also specifically in the context of India by Joshi et al. (2011) and Roychowdhury, Jones, and Arrowsmith (2009). The conclusion of these studies is that none of these sources is error-free, and that there is no strong case for preferring day-time over night-time satellite imagery if aerial photographs are not systematically available for the area to be mapped.

[4] DMSP-OLS night-time imagery overestimates the actual extent of lit area on the ground, due to a combination of coarse spatial resolution, overlap between pixels, and minor geolocation errors (Small, Pozzi, and Elvidge 2005).

things control for these differences in data sources, as well as for different calibrations of the night-time lights satellites.

The resulting panel dataset of urban footprints is unbalanced for two reasons: first, some settlements become large enough to be detectable only later in the panel; second, some settlements appear as individual cities for some years in the panel, and then become part of larger urban agglomerations in later years. The number of cities in the panel ranges from 351 to 457, depending on the year considered. As a result, the "long-difference" sample used in the baseline specifications includes 351 observations. In Appendix Table A4, I show that the results continue to hold in the full sample.

The criterion for being included in the analysis is to appear as a contiguous lighted shape in the night-time lights dataset. This appears to leave out only very small settlements.

A2. Shape metrics

The notion of "compactness" of an urban footprint is borrowed from the urban planning and landscape ecology literature (Angel, Civco, and Parent 2009, 2010). Intuitively, the geometric concept of compactness rests on the idea that the circle is the most compact shape. The extent to which a polygon's shape departs from that of a circle can be measured through many distinct indexes, all based on the distribution of points within a polygon.

My benchmark indicator throughout the paper is the disconnection index (corresponding to the "cohesion" index in Angel, Civco, and Parent (2010)). It is defined as the average Euclidean distance between all pairs of interior points within a polygon and can be viewed as a proxy for the length of all potential trips within the city, without restricting one's attention to those to or from the center. Higher values of the index denote longer distances within the city and less compact shape. Specifically, consider n random interior points sampled within a polygon and denote the distance between points j and i as d_{ij}. The index is defined as follows:

$$S = \frac{\sum_{i=1}^{n} \sum_{j}^{n} d_{ij}}{n(n-1)}.$$

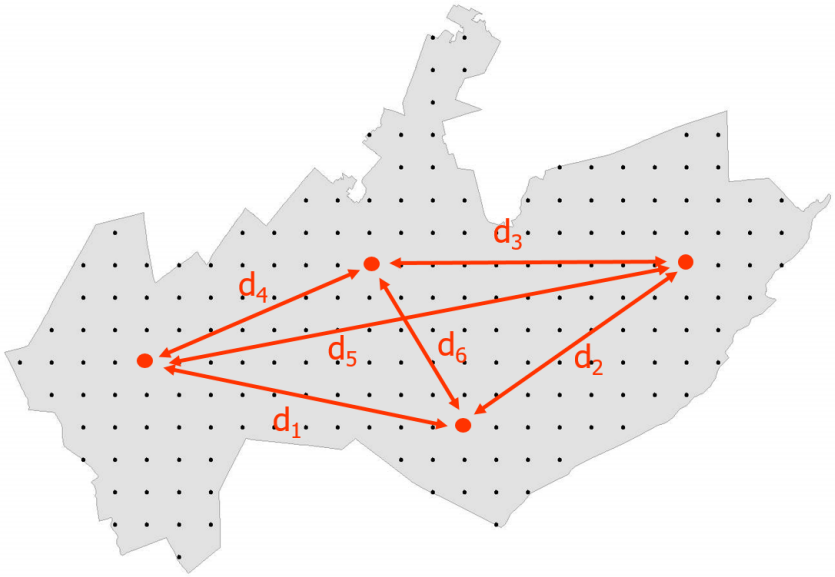
This is illustrated in Figure A.1 for four hypothetical sample points.

I compute the index using the Angel, Civco, and Parent (2009) ArcGIS routines.[5] The Shape Metrics tool considers 20,000 interior points, uniformly distributed throughout the polygon in a grid pattern. Then, for computational ease, the index is computed for 30 samples of 1000 randomly selected points within this set and averaged. I compute the index in kilometers.

The disconnection index is mechanically correlated with polygon area. In order to disentangle the effect of geometry *per se* from that of city size, two approaches are possible. One is to explicitly control for the area of the footprint, which I do in the baseline specification throughout the paper. Alternatively, the index can be normalized, computing a version that is invariant to the area of the polygon. I do so by computing first the radius of the "Equivalent Area Circle" (EAC), namely a circle with an area equal to that of the polygon. I then normalize the index of interest by dividing it by the EAC radius.[6]

[5] I am thankful to Vit Paszto for help with the ArcGIS shape metrics tool.

[6] My normalization is slightly different from that proposed by Angel, Civco, and Parent (2009).



$$S = \frac{d_1 + d_2 + \cdots + d_6}{6}$$

Figure A.1: Calculation of the disconnection index, example from Angel, Civco, and Parent (2009).

Figure A2 reports the disconnection index computed for selected shapes, where S represents the non-normalized index and nS the normalized version. These examples illustrate departures from the circular shape that are associated with higher values of the disconnection index. Elongated shapes (such as v in the figure) and polygons with recesses and gaps (such as iii and iv, similar to urban areas growing around topographic obstacles) are all associated with greater disconnection.

The shortest connecting paths used in the computation of average distances do not need to lie within the polygon. In this regard, the index may underestimate distances that account for the placement of roads. Furthermore, the index is defined for contiguous polygons only: as a result, a built-up area disconnected from the main urban footprint would not contribute to the index. In this respect, the index may underestimate the disconnectedness of non-contiguous development.

For robustness, I also compute three additional shape metrics:

(i) The *remoteness* index ("proximity" in Angel, Civco, and Parent (2010)) is the average distance between all interior points and the centroid.[7] It can be viewed as a proxy for the average length of all potential trips to the center.

(ii) The *spin* index is computed as the average of the squared distances between interior points and the centroid. This is similar to the remoteness index, but gives more weight to the polygon's extremities, corresponding to the periphery of the footprint. This index has particularly high values for

[7]The centroid of a polygon, or center of gravity, is the point that minimizes the sum of squared Euclidean distances between itself and each vertex.

footprints that have tendril-like projections.

(iii) The *range* index captures the maximum distance between two points on the shape perimeter, representing the longest possible distance between two points within the city.

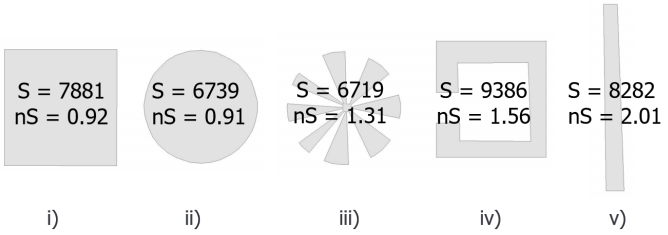


Figure A.2: Disconnection index for a sample of polygons, adapted from Angel, Civco, and Parent (2009). S denotes the disconnection index, and nS the normalized version.

Similarly to the benchmark indicator, all the indexes above are measured in kilometers, higher values denoting a greater departure from circularity and longer within-city distances, and can be normalized by the EAC radius.

Even though these indexes represent independent properties, in practice they tend to be highly correlated and should be viewed as different proxies for the same broad notion of "compactness". In my dataset, the correlation between any two indexes is between 0.82 and 0.97 (except for the correlation between spin and range that is 0.7). Table A6 and A7 in the Appendix shows that the first-stage and IV results are robust to employing different indexes.

A3. Outcome variables

Population City-level data for India is difficult to obtain. The only systematic source that collects data explicitly at the city level is the Census of India, conducted every 10 years. I employ population data from Census years 1871-2011. As explained in Section 5, historical population (1871-1941) is used to construct the instrument, whereas population drawn from more recent waves (1951, 1991, 2001, and 2011) is used as an outcome variable. Specifically, I combine the following sources:

- 1871-1951 population data harmonized by Mitra (1980).
- Population totals by urban agglomerations 1981-2011 (Office of the Registrar General & Census Commissioner, India 1981-2011), assembled by CityPopulation.de.
- Primary Census Abstract 1991 (Office of the Registrar General & Census Commissioner, India 1991a), accessed in CD-ROM format through Harvard University Libraries.
- Primary Census Abstract 2001 (Office of the Registrar General & Census Commissioner, India 2001a), accessed in CD-ROM format through Harvard University Libraries.
- H-series tables 2001 (Office of the Registrar General & Census Commissioner, India 2001b), accessed in CD-ROM format through Harvard University Libraries.

- Code List for Land Regions 2001 (State, District, Sub-District, Town, Village) (Office of the Registrar General & Census Commissioner, India 2001c), accessed through the Census website.
- Village-level amenities data 2001 (Office of the Registrar General & Census Commissioner, India 2001d), accessed in CD-ROM format through Harvard University Libraries.
- Village and town location codes 2001 (Office of the Registrar General & Census Commissioner, India 2001e), accessed in CD-ROM format through Harvard University Libraries.
- Tables 3 and 6 2001 (Office of the Registrar General & Census Commissioner, India 2001f and 2001g), accessed through the Census website.
- Primary Census Abstract 2011 (Office of the Registrar General & Census Commissioner, India 2011a), accessed through the Census website.
- List of Villages/Towns 2011 (Office of the Registrar General & Census Commissioner, India 2011b), accessed through the Census website.

Note that "footprints", as retrieved from the night-time lights dataset, do not always have an immediate Census counterpart in terms of town or urban agglomeration, as they sometimes stretch to include suburbs and towns treated as separate units by the Census. This is especially problematic for earlier Census waves. A paradigmatic example is the Delhi conurbation, which as seen from the satellite expands well beyond the administrative boundaries of the New Delhi National Capital Region. When assigning population totals to an urban footprint, I sum the population of all Census settlements that are located within the footprint, thus computing a "footprint" population total. Moreover, in order to assemble a consistent panel of city population totals over the years one also has to account for changes in the definitions of "cities", "urban agglomerations", and "outgrowths" across the earlier Census waves. Mitra (1980) provides harmonized figures for all Census waves up to 1971 and I addressed this issue in subsequent waves.

Wages and rents For wages and rents, I rely on the National Sample Survey and the Annual Survey of Industries, which provide, at most, district identifiers (National Sample Survey Office 2007-2008; Central Statistics Office 2009-2010). I thus follow the approach of Greenstone and Hanna (2014) and Chauvin et al. (2017): I match cities to districts and use district urban averages as proxies for city-level averages. It should be noted that the matching is not always perfect, for a number of reasons. First, it is not always possible to match districts as reported in these sources to Census districts, and through these to cities, due to redistricting and inconsistent numbering throughout this period. Second, there are a few cases of large cities that cut across districts (e.g., Hyderabad). Finally, there are a number of districts which contain more than one city from my sample. I provide results for three samples: one including any city that can be matched; one that only includes cities for which there is a one to one mapping with a district; and finally a sample where I only include the top city in each district. The matching process introduces considerable noise and leads to results that are relatively less precise and

less robust than those I obtain with city-level outcomes.[8] Wages and rents are converted to current rupees using a GDP deflator (International Monetary Fund 1970-).

Data on wages are taken from the Annual Survey of Industries (Central Statistics Office 1990-1991, 2009-2010).[9] [10] These are repeated cross-sections of plant-level data collected by the Ministry of Programme Planning and Implementation of the Government of India. The ASI covers all registered manufacturing plants in India with more than fifty workers (one hundred if without power) and a random one-third sample of registered plants with more than ten workers (twenty if without power) but less than fifty (or one hundred) workers. As mentioned by Fernandes and Sharma (2012) amongst others, the ASI data are extremely noisy in some years, which introduces a further source of measurement error. In the main long difference specifications, I employ the 1990 and 2010 waves. Results are similar using intermediate waves in a panel specification. Data from intermediate waves 1994, 1995, 1998 are also employed for the robustness check in Table A14.

A drawback of the ASI data is that it covers the formal manufacturing sector only.[11] This may affect the interpretation of my results, to the extent that this sector is systematically over- or underrepresented in cities with worse shapes. I examined the relationship between shape and the industry mix of cities, employing Economic Census data, and found no obvious patterns. The share of manufacturing appears to be slightly lower in non-compact cities, but this figure is not significantly different from zero, which somewhat alleviates the selection concern discussed above.

Unfortunately, there is no systematic source of data for property prices in India across a sufficient number of cities. I construct a proxy for the rental price of housing drawing on the National Sample Survey (Household Consumer Expenditure schedule), which asks households about the amount spent on rent (National Sample Survey Office 2005-2006, 2007-2008).[12] In the case of owned houses, an imputed figure is provided . Rounds 62 (2005-2006), 63 (2006-2007), and 64 (2007-2008) are the only ones for which the urban data is representative at the district level and which report total dwelling floor area as well. In my baseline specification I focus on rounds 62 and 64 and take a long difference, but results are similar using all three waves in the panel version.

I construct a measure of rent per square meter based on total rent amount and floor area. These

[8] To facilitate the harmonization of districts across Census waves, I draw upon the 2011 Administrative Atlas of India, available from the Census website (Office of the Registrar General & Census Commissioner, India 2011c; Kumar and Somanathan 2009).

[9] The data is confidential, but may be obtained with Data Use Agreements with the Indian Ministry of Statistics and Programme Implementation. Researchers interested in accessing the data must obtain CD-ROMS which can be obtained by contacting Deputy Director General, Computer Centre, East Block-10, R K Puram, New Delhi at ddg.cc-mospi@gov.in. Also see: http://mospi.gov.in/support-queries.

[10] To facilitate the matching of ASI districts to Census districts, I draw upon the ASI directories provided by Adhvaryu, Chari, and Sharma (2013).

[11] An alternative source of wages data is the National Sample Survey, Employment and Unemployment schedule. This provides individual-level data that cover both formal and informal sector. However, it is problematic to match these data to cities. For most waves, the data are representative at the NSS region level, which typically encompasses multiple districts.

[12] The data is confidential, but may be obtained with Data Use Agreements with the Indian Ministry of Statistics and Programme Implementation. Researchers interested in accessing the data must obtain CD-ROMS which can be obtained from the Deputy Director General, Computer Center, M/O Statistics and PI, East Block No. 10 R.K. Puram, New Delhi-110066 by remitting the price along with packaging and postal charges. Also see: http://mospi.gov.in/support-queries.

figures are likely to be underestimating the market rental rate, due to the presence of rent control provisions in most major cities of India (Dev 2006). While I cannot observe which figures refer to rent-controlled housing, as an attempt to cope with this problem, I also construct an alternative proxy for housing rents which focuses on the upper half of the distribution of rents per meter. This is *a priori* less likely to include observations from rent-controlled housing.

For robustness, I also consider an alternative source, used by Chauvin et al. (2017): the India Human Development Survey (IHDS), comprising two rounds (2005 and 2012) (Desai et al. 2005, 2012). The IHDS is a nationally representative household survey including 971 urban neighborhoods across India. It reports monthly expenditures on housing rents as well as housing characteristics (other than dwelling size), including: number of rooms, house type (house with no shared walls, house with shared walls, flat, chawl, slum housing, or other), housing surrounded by sewage, predominant wall type (grass/thatch, mud/unburnt bricks, plastic, wood, burned bricks, GI sheets or other metal, stone, or cement/concrete), predominant roof type (grass/thatch/mud/wood, tile, slate, plastic, GI metal/asbestos, cement, brick, or stone concrete), and predominant floor type (mud, wood/bamboo, brick, stone, cement, tiles/mosaic, or others). This is an advantage over the NSS data as I can run a hedonic regression of rents on the above characteristics and consider the residuals. At most, district identifiers are provided, and not all districts can be matched to Census ones. As a result, only about 260 cities can be matched to an IHDS district.

A4. Correlates of city shape and controls

Below I discuss the results and the sources of the city-level variables employed in Table A2 in this Appendix. Many of these variables are also employed throughout the paper in a number of robustness checks. Summary statistics are provided in Table A3. Unless otherwise specified, all the distance variables are in kilometers and are calculated from the Central Business District (CBD) of each city (based on the centroid of the 1950 footprint).

Table A2 shows the correlation between city shape, in levels and changes, and a number of city-level attributes. Each row presents the OLS coefficients of two distinct regressions: in column 1 I regress city shape in 2010 on the relevant city attribute, controlling for 1950 city area; column 2 is similar, but the dependent variable is the 1950-2010 long difference in city shape and I control for city area in 1950.

In Panel A, I rank cities by their 1951 population and assign quartile dummies to each. The bottom three quartiles are all associated with more compact shapes whereas the top quartile is associated with less compact shape. This pattern holds both in levels (column 1) and in changes (column 2). In Section 5 I argue that the correlation between bad shape and city size is spurious and driven by the tendency of cities to deteriorate in shape as they expand.

In Panel B, I consider the channels highlighted by the urban planning literature: access to public services and urban transit. Specifically, I consider the share of households with connections to electricity, with tap water on premises, and with cars, from the 2011 houselisting and housing Census tables (Office of the Registrar General & Census Commissioner, India 2011d) which are accessible from the

census website.[13] Less compact cities are associated with a larger share of households connected to public services. This runs counter the predictions of the urban planning literature, arguing that compact cities provide better service access. However, any potential causal effects are likely to be confounded by the fact that less compact cities tend to be the largest and highest-income cities in the sample. This can also explain the correlation between non-compact shape and the share of households with cars.

Next, I consider the length of the urban road network. I overlap 2019 road maps from OpenStreetMap (OpenStreetMap contributors 2019) on 2010 city outlines as defined by the nightlights, following the approach of Akbar et al. (2019). OpenStreetMap is a collaborative worldwide mapping project. A caveat in employing these data is that the degree of accuracy and comprehensiveness may vary across cities, raising concerns of measurement error. Despite disconnected cities being larger and more developed, they do not appear to have a denser road network. In levels, higher values of the disconnection index are associated with a shorter urban road network. In changes, the sign becomes negative, but the effect size is small: holding city area constant, as the average within-city trip increases by one kilometer, the road network expands by one meter. This suggests that the provision of infrastructure in non-compact cities may indeed be more difficult, as urban planners suggest.

Interestingly, non-compact shape is positively correlated, both in levels and in changes, with the average distance to workplace in 2011. This variable is calculated based on district-level Census data, accessible from the Census website (Table B-28), on the number of urban workers residing at different reported distances from their workplaces, by coarse bins (0-1, 2-5, 6-10, 11-20, 21-30, 31-50, or above 50 kilometers) (Office of the Registrar General & Census Commissioner, India 2011e). I calculate a district-level average distance to work by averaging the mean distance within each bin,[14] weighted by the share of workers. I then match each city to a district.[15] This should be viewed as a noisy proxy for commuting distance within the city, as the distance bins are coarse and include large distances relative to the average city size. The correlation with city shape (in levels) is positive, suggesting that less compact cities may be associated with longer commutes. Furthermore, the positive correlation with changes in shape indicates that cities with long commutes are cities that became less compact than they were. This is plausible for a city that starts out as monocentric and compact, and grows into a less compact shape over time, with commutes to the center becoming longer over time. Conversely, commutes may not be as long in a city that has always been non-compact and perhaps more polycentric. These patterns are explored further in Section 9 in the paper.

In Panel C, I consider pre-determined city characteristics related to geography and geology:

- Elevation is measured in the CBD, based on data from the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (NASA and METI 2011). Mountainous cities are defined as having elevation above 700 meters.

[13] A number of variables employed in the paper are drawn from the Census town-level tables. It should be noted that these tables cover Census towns only, excluding small settlements that may fall within a city's lit-up footprint.

[14] I consider 60 kilometers for the "above 50 kilometers" bin.

[15] I show results for all cities, but results are similar when excluding districts with multiple cities or when considering only the main city in each district.

- Distance from the coast is based on the Global Self-consistent Hierarchical High-resolution Shorelines (GSHHS) dataset (Wessel and Smith, 2013).

- Distance from the nearest river or lake is measured combining large rivers from the Natural Earth 2.0.0 dataset (Patterson and Kelso 2012) and lakes from the WWF Global Lakes and Wetlands Database, Level 2 (World Wildlife Fund and the Center for Environmental Systems Research 2004). The river/lake dummy is equal to 1 for cities whose CBD lies within 5 kilometers of a river or lake.

- Distance from nearest mineral deposit is calculated based on the location of mineral deposits recorded in the Mineral Resources Data System dataset, assembled by the U.S. Geological Survey (U.S. Geological Survey 2005). A city is considered with a mineral deposit if there is a deposit within 50 kilometers of the CBD.

- Ruggedness (in meters) is drawn from the G-Econ gridded dataset (Chen and Nordhaus 2016) and is measured at the 1 degree (approximately 100 kilometers) level. It is calculated based on the average absolute change in elevation between adjacent 10 Arc-minutes cells included in each 1 degree cell. Higher values imply more variation in elevation and greater terrain ruggedness. I match each city CBD to the corresponding 100-km grid cell and assign the corresponding value.

- Bedrock depth (in meters) is drawn from the SoilGrids dataset (Hengl et al. 2014), a global gridded dataset at a 1 km resolution. I take the average bedrock depth within 100 kilometers of a city's CBD.

- Crop suitability (in tonnes per hectare per year) is calculated based on the potential yields (for low-input, rainfed production) of the top 5 most suitable crops in India (dryland rice, wetland rice, maize, millet, sorghum), drawn from the FAO's Global Agro-Ecological Zones (GAEZ) dataset (FAO/IIASA 2010). The raw data is available at a resolution of 5 arc minutes (approximately 10 kilometers). Yields are averaged within 100 kilometers of the CBD.

Elevation, distance from the coast, distance from water bodies, and terrain ruggedness capture the presence of geographic constraints to city expansion. However, the raw correlation with city shape is insignificant, except for ruggedness which is associated with less compact cities. This suggests that what affects city shape is not the generic presence of particular geographic features and one may have to account for the exact position of geographic constraints - which motivates the way in which I construct my instrument. Similarly, I find no meaningful correlation with bedrock depth, that has been associated in the literature with higher construction costs for high-rises (Barr, Tassier, and Trendafilov 2011). Crop suitability and the distance from mineral deposits, which may affect the city's productivity, are also not significant.

Finally, in Panel D I consider other, non-predetermined city characteristics. The British direct rule dummy identifies cities in districts that were formerly part of British India, based on Iyer (2010). Distance from state headquarters, from district headquarters, and from the nearest city with more than

100,000 inhabitants are drawn from the 2011 Census Town Amenities tables (Office of the Registrar General & Census Commissioner, India 2011f), available from the Census website. As expected, shape is persistent in time, as highlighted by the positive coefficient for initial shape. More remote cities, further away from state or district headquarters, tend to be more compact, but the correlation is weak. Cities that were under direct British rule are on average less compact, consistent with the findings of Baruah, Henderson, and Peng (2017) on British colonial cities being more sprawled, but this correlation is only borderline significant. Conversely, there is a strong tendency of state capitals to deteriorate in shape, probably because they are also the largest cities.

A5. Infrastructure and transit

Below I discuss the data related to transit and infrastructure employed in Tables 9, A15, and A16.[16]

Current road network I measure the current length of the road network by overlapping digital roadmaps from 2019 OpenStreetMap (OpenStreetMap contributors 2019) with 2010 city outlines as defined from the night-time lights. In Table A16 I also consider motorways, defined as the subset of road segments labeled by OpenStreetMap as "motorways" or "trunks", corresponding to dual-carriage roads similar to freeways in the U.S. (Akbar et al. 2019).

Indices by Akbar et al. (2018, 2019) In Tables 9, A15, and A16 I consider indices developed for Indian cities by Akbar et al. (2018, 2019).[17] The authors provide estimates of the unit cost of commuting in Indian cities using transit times predicted by Google Maps in 2016, and aggregate these estimates into city-level indices of vehicular mobility. Their methodology is primarily based on feeding into Google Maps origin-destination pairs and collecting information on the duration and length of these artificial trips. In addition, they also create indices related to the spatial properties of the road network in a city, defined overlapping 2016 OpenStreetMap with city outlines defined using a combination of night-time lights and other satellite-based products.

The proximity index (Akbar et al. 2018) is a measure of distance-based accessibility. It is based on the road distance between random points in the city and a number of amenities (shopping centers, train stations etc.), selected by Google Maps within a pre-specified radius. Higher values indicate greater accessibility and shorter trips.

The grid conformity index (Akbar et al. 2019) measures the extent to which the city's 2016 road network is laid out as a regular grid. It measures the proportion of edges in a city's road network that conform to the dominant grid orientation, by being perpendicular or parallel to the modal edge bearing. Higher values indicate more regular grids and correlate with better vehicular mobility.

The mobility index (Akbar et al. 2019) is their benchmark index of vehicular mobility and is based on the speed of simulated trips. This index abstracts from city shape as the length of the simulated trips is pre-specified by the authors. Factors affecting this index include road density, road quality, and traffic congestion.

[16] The data on work commutes employed in Table 10 is discussed in Section A4 above.

[17] I am thankful to the authors for granting me access to their data.

Note that, *a priori*, there is no clear mapping between city shape and the Akbar et al. (2018, 2019) indices, as the latter depend on the internal functioning of the road network and traffic patterns, and not on the city's layout. In Table A16, panel B, I provide IV and OLS estimates for the relationship between city shape and the three indices discussed above, subject to the caveat of very weak instruments. Poor city shape is associated with lower mobility, both in the OLS and the IV. This may stem from disconnected cities having a less functional road network, as highlighted in Table A16, panel A. However, the magnitudes are small: according to the most conservative point estimate, for a one standard deviation deterioration in city shape, mobility declines by 1.2%. Results are more mixed when considering proximity. In the IV, bad shape is associated with lower proximity, but the coefficient is small and insignificant at conventional levels. The corresponding OLS is positive and significant, perhaps reflecting the fact that in the OLS bad shape correlates with city size, and large cities tend to have more amenities to begin with.

Interaction variables in Tables 9 and A15 Below I discuss the variables employed in the interaction specifications of Table 9 and A15. These tables present the same IV specifications, but Table A15 additionally controls for a proxy for city income (the number of banks in 1981) from the Census Town Directory (Office of the Registrar General & Census Commissioner, India 1981).

In columns 1 through 3 I interact city shape with urban road length. In column 1 I consider the total length of urban roads in a city, obtained from OpenStreetMap as discussed above. In column 2 I consider the total length of city urban roads, from the 1981 Census Town Directory (Office of the Registrar General & Census Commissioner, India 1981). In column 3 I consider the total length of urban roads in a state as of 1981, from the Ministry of Road Transport and Highways (Transport Research Wing 1971-2020a), accessed through the Centre for Monitoring Indian Economy's (CMIE) website. This figure is normalized by the total urban land area in a state, in square kilometers, as provided by the Centre for Industrial and Economic Research's Industrial Databooks (CIER 1990).

In columns 4 and 5 I consider the proximity and grid conformity index from Akbar et al. (2018, 2019), discussed above.

In columns 6 through 8 I interact city shape with proxies for the availability of motor vehicles. In column 6 (7) I consider the number of city households with access to cars, reported in the 2011 (2001) Census. Specifically, the 2001 figure is drawn from Table H-13 (Office of the Registrar General & Census Commissioner, India 2001h) and the 2011 figure is drawn from Table Hh-14 (Office of the Registrar General & Census Commissioner, India 2011d), both accessible from the Census website. In column 8 I consider the total number of registered cars in a state in 1984, from the Ministry of Road Transport and Highways (Transport Research Wing 1971-2020b), accessed through the Centre for Monitoring India Economy. Year 1984 is the earliest year for which this figure is available for most states. This figure is normalized by the total urban land area in a state in 1981, in square kilometers, as provided by the Centre for Industrial and Economic Research's Industrial Databooks (CIER 1990).

A6. Other variables

Geographic constraints For the purposes of constructing the city shape instrument, I code geographic constraints to urban expansion as follows. Following Saiz (2010), I consider land pixels as "undevelopable" when they are either occupied by a water body, or characterized by a slope above 15%. I draw upon two high-resolution sources: the Advanced Spaceborne Thermal Emission and Reflection Radiometer (ASTER) Global Digital Elevation Model (NASA and METI 2011), with a resolution of 30 meters, and the Global MODIS Raster Water Mask (Carroll et al. 2009), with a resolution of 250 meters. I combine these two raster datasets to classify pixels as "developable" or "undevelopable". Figure 4 in the paper illustrates this classification for the Mumbai area.

Firm location and employment subcenters Data on the spatial distribution of employment in year 2005 is derived from the urban Directories of Establishments, pertaining to the 5th Economic Census (Office of the Registrar General, India 2005). For this round, establishments with more than 10 employees were required to provide an additional "address slip", containing a complete address of the establishment, year of initial operation, and employment class. I geo-referenced all the addresses corresponding to cities in my sample through Google Maps API, retrieving consistent coordinates for approximately 240 thousand establishments in about 190 footprints.[18]

I utilize these data to compute the number of employment subcenters in each city, following the two-stage, non-parametric approach described in McMillen (2001). Of the various methodologies proposed in the literature, this is particularly suitable for my context as it can be fully automated and replicated for a large number of cities. This procedure identifies employment subcenters as locations that have significantly larger employment density than nearby ones, and that have a significant impact on the overall employment density function in a city. This procedure is outlined in Section F of this Appendix. The number of employment subcenters calculated for year 2005 ranges from 1, for purely monocentric cities, to 9, for large cities such as Delhi and Mumbai. Consistent with results obtained in the U.S. context by McMillen and Smith (2003), larger cities tend to have more employment subcenters.

Slum population Data on slums is drawn from the 1981 and 2011 Census, which provide slum population counts for selected cities. Specifically, the 1981 data is drawn from the Ministry of Urban Affairs (1981-2001), accessed through the Indiastat website. The 2011 data is drawn from the Primary Census Abstract for Slums, Town level (Office of the Registrar General & Census Commissioner, India 2011g), accessed through the Census website.

The Census defines slums as follows: all areas notified as "slum" by state or local Government; and any compact area with population above 300 characterized by "poorly built congested tenements, in unhygienic environment, usually with inadequate infrastructure and lacking in proper sanitary and drinking water facilities". Such areas are identified by Census Operations staff.

[18] Results are similar to excluding firms whose address can only be approximately located by Google Maps.

Floor Area Ratios Data on the maximum permitted Floor Area Ratios for a small cross-section of Indian cities (55 cities in my sample)[19] is taken from Sridhar (2010), who collected them from individual urban local bodies as of the mid-2000s. FARs are expressed as ratios of the total floor area of a building over the area of the plot on which it sits. I consider the average of residential and non-residential FARs (but results are similar focusing on residential FARs only). For a detailed discussion of FARs in India, see Sridhar (2010) and Bertaud and Brueckner (2005).

While in this paper I take FARs as given, the question might arise on their determinants. Regressing FAR values on urban shape and area, I found weak evidence of FARs being more restrictive in larger cities, consistent with one of the stated objectives of regulators - curbing densities in growing cities.

Electrical connections and tap water Data on the availability of electrical connections and tap water on premises is drawn from the 1991 and 2011 Census. Specifically, the 1991 figure is drawn from the Primary Census Abstract H series tables (Office of the Registrar General & Census Commissioner, India 1991b), accessed in CD-ROM format through Harvard University Libraries. The 2011 figure is drawn from Tables Hh-14 (Office of the Registrar General & Census Commissioner, India 2011d) and Tables HH-11 (Office of the Registrar General & Census Commissioner, India 2011h), accessed through the Census website.

B. A Simple Model of Spatial Equilibrium across Cities

I motivate the empirical analysis of the impacts of city shape drawing on a model of spatial equilibrium across cities (Rosen 1979; Roback 1982). I embed city shape in this framework by hypothesizing that households and firms may value city compactness when evaluating the trade-offs associated with different cities. In order to deliver the intuition and provide estimable equations, I focus on a simple version of the model, with Cobb-Douglas functional forms, following the exposition in Glaeser (2008). I then discuss caveats and extensions to be addressed in future research.

Model setup

The model features homogeneous households, firms, and developers.

Households have Cobb-Douglas utility $U(C,H,\theta) = \theta C^{1-\alpha} H^{\alpha}$ over a numéraire good C, housing H, and a city-specific "quality of life" parameter θ. The latter captures any utility cost or benefit associated with living in a particular city that requires compensation through factor prices. It is useful to conceptualize θ as consisting of three components: "public services" θ_P, "transit accessibility" θ_T, and "consumption amenities" θ_A. All else being equal, better public services (such as electricity or water), greater accessibility, and better amenities (such as good climate) improve household utility. Denoting city shape with S, I assume that S can affect θ_P and θ_T, in line with the conjectures of urban planners. Households supply labor inelastically, receiving a city-specific wage W. Solving their

[19] Sridhar (2010) collects data for about 100 cities, but many of those cities are part of larger urban agglomerations, and do not appear as individual footprints in my panel, or are too small to be detected by night-time lights.

utility-maximization problem, for a given city, yields the following indirect utility:

$$\log(W) - \alpha \log(p_h) + \log(\theta) = \log(\overline{v}). \tag{B.1}$$

where p_h is the rental price of housing.

Spatial equilibrium requires that indirect utility $\overline{v}$ be equalized across cities, otherwise workers would move.[20] This condition delivers the key intuition that households, in equilibrium, implicitly pay for better quality of life, as captured by θ, through lower wages (W) or through higher housing prices (p_h).

In the production sector, firms competitively produce a traded good Y, using labor N, traded capital K (which trades at price 1), a fixed supply of non-traded capital $\overline{Z}$[21], and a bundle of city-specific production amenities A. Their production function is $Y(N,K,\overline{Z},A) = AN^{\beta}K^{\gamma}\overline{Z}^{1-\beta-\gamma}$. Similar to households, firms may benefit from compact city shape through better access to services or because of greater accessibility, which I capture by allowing S to affect A via two components, A_P and A_T. Normalizing the price of traded capital to 1, the zero-profit condition for firms delivers the following labor demand curve:

$$(1-\gamma)\log(W) = (1-\beta-\gamma)(\log(\overline{Z}) - \log(N)) + \log(A) + \kappa_1. \tag{B.2}$$

Finally, developers competitively produce housing H, using land l and "building height" h. In each city there is a fixed supply of land $\overline{L}$, determined by planners.[22] Denoting the price of land with p_l, the developers' cost function reads $C(H) = c_0 h^{\delta} l - p_l l$, with $\delta > 1$.

By combining housing supply, obtained from the developers' maximization problem, with housing demand, resulting from the households' problem, one obtains the following housing market equilibrium condition:

$$(\delta-1)\log(H) = \log(p_h) - \log(c_0\delta) - (\delta-1)\log(N) + (\delta-1)\log(\overline{L}). \tag{B.3}$$

Equilibrium

The system of equations given by the three optimality conditions (B.1), (B.2), and (B.3) can be solved for the three endogenous variables N, W, and p_h as functions of the city-specific productivity parameter A and consumption amenities θ. Denoting with F, G, D, and K constant functions of the model's deep parameters, this yields the following:

$$\log(N) = F_N \log(A) + G_N \log(\theta) + D_N \log(\overline{L}) + K_N \tag{B.4}$$

[20] The notion of spatial equilibrium across cities presumes that households are choosing across various locations. While mobility in India is lower than in other developing countries (Munshi and Rosenzweig 2016), the observed pattern of migration to urban areas is compatible with this element of choice: as per the 2001 Census, about 38% of rural to urban internal migrants move to a location outside their district of origin, presumably choosing a city rather than simply moving to the closest available urban area.

[21] This is to ensure decreasing returns at the city level, which, in turn, is required to have a finite city size. Alternatively, one could assume congestion in amenities or decreasing returns in housing production.

[22] In this framework, the amount of land to be developed is assumed to be given in the short run. It can be argued that, in reality, this is an endogenous outcome of factors such as regulation, city growth, and geographic constraints. In my empirical analysis I incorporate city area as a control variable and I instrument it using historical population, thus abstracting from these issues.

$$\log(W) = F_W \log(A) + G_W \log(\theta) + D_W \log(\overline{L}) + K_W \tag{B.5}$$

$$\log(p_h) = F_P \log(A) + G_P \log(\theta) + D_P \log(\overline{L}) + K_P \tag{B.6}$$

where $F_N, F_W, F_P > 0$; $G_N, G_P > 0$; and $G_W < 0$.

Population, wages, and rents are all increasing functions of the city-specific productivity parameter A. Intuitively, higher A allows firms to pay higher wages, which attracts households and bids up rents. Similarly, population and rents are increasing in the "quality of life" parameter θ: better amenities attract households and bid up rents. Wages are decreasing in θ because firms prefer cities with higher production amenities, whereas households prefer cities with higher consumption amenities, and factor prices - W and p_h - strike the balance between these conflicting location preferences.

Reduced-form predictions

Consider now an exogenous shifter of urban geometry S, higher values denoting less compact shapes. I hypothesize that S may be part of the A or θ bundle as follows:

$$\log(A) = \kappa_A + \lambda_A S \tag{B.7}$$

$$\log(\theta) = \kappa_\theta + \lambda_\theta S. \tag{B.8}$$

Plugging (B.7) and (B.8) into (B.4), (B.5), (B.6) yields the following reduced-form equations:

$$\log(N) = B_N S + D_N \log(\overline{L}) + K_N \tag{B.9}$$

$$\log(W) = B_W S + D_W \log(\overline{L}) + K_W \tag{B.10}$$

$$\log(p_h) = B_P S + D_P \log(\overline{L}) + K_P. \tag{B.11}$$

Suppose that non-compact shape S decreases households' indirect utility in equilibrium, but does not directly affect firms' productivity (λ_A=0, λ_θ<0). This would be the case if, for example, households located in non-compact cities faced longer commutes, or were forced to live in a less preferable location so as to avoid long commutes, while firms' transportation costs were unaffected - possibly because of better access to transportation technology, or because of being centrally located within a city. In this case, the model predicts that $B_N < 0$, $B_W > 0$, $B_P < 0$. A city with poorer shape should have, all else equal, smaller population, higher wages, and lower housing rents. Intuitively, households prefer cities with good shapes, which drives rents up and bids wages down in these locations.

Suppose, instead, that poor city geometry enters both θ and A, i.e. in equilibrium it is associated with both lower household indirect utility and lower firm productivity (λ_A<0, λ_θ<0). This would be the case if the costs of longer commutes were borne by households but also by firms. This would imply $B_N < 0$, $B_W \gtrless 0$, $B_P < 0$. The model's predictions are similar, except that the effect on wages will be ambiguous, given that now both firms and households prefer to locate in compact cities. As both firms and households compete to locate in compact cities, the net effect on W depends on whether firms or households value low S relatively more (on the margin). If S affects households more than firms, then $B_W > 0$.

Denote with $\widehat{B_N}$, $\widehat{B_W}$, and $\widehat{B_P}$ the reduced-form estimates for the impact of S on, respectively, $\log(N), \log(W)$, and $\log(p_h)$. These estimates, in conjunction with plausible values for parameters α, β, and γ, can be used to back out λ_θ and λ_A, representing respectively the marginal willingness to pay for S and the marginal productivity impact of S. Totally differentiating the indirect utility of households (B.1) with respect to S yields:

$$\frac{\partial \log(\theta)}{\partial S} = \alpha \frac{\partial \log(p_h)}{\partial S} - \frac{\partial \log(W)}{\partial S} \tag{B.12}$$

suggesting that λ_θ can be estimated as

$$\widehat{\lambda_\theta} = \alpha \widehat{B_P} - \widehat{B_W}. \tag{B.13}$$

Totally differentiating the zero-profit condition (B.2) with respect to S yields:

$$\frac{\partial \log(A)}{\partial S} = (1-\beta-\gamma)\frac{\partial \log(N)}{\partial S} + (1-\gamma)\frac{\partial \log(W)}{\partial S} \tag{B.14}$$

suggesting that λ_A can be estimated as

$$\widehat{\lambda_A} = (1-\beta-\gamma)\widehat{B_N} + (1-\gamma)\widehat{B_W}. \tag{B.15}$$

Equations (B.9), (B.10), and (B.11) are taken to the data in Section 6 in the paper. Estimates of λ_A and λ_θ are provided in Section 8.

Discussion and extensions

The framework outlined above makes a number of simplifying assumptions and modelling choices. In what follows I discuss limitations of the current framework and extensions for future research.

The model features homogeneous and perfectly mobile households. As such, it has little to say about welfare consequences of bad shape, as all agents are marginal and indifferent in equilibrium. With indirect utility pinned down by utility in a reservation location, there are no welfare gains from improving shape, as higher rents accrue to landlords. In order to be able to make welfare and distributional statements, one would require a richer model incorporating landlords and tenants as well as heterogeneity in idiosyncratic location preferences or migration costs.

However, the willingness to pay parameter λ_θ can be viewed as an upper bound for welfare effects of deteriorating shape, to the extent that reality is somewhere in between the case with infinitely elastic or infinitely inelastic supply of urban dwellers. With a fixed total urban population at the country level, equilibrium indirect utility $\overline{v}$ will unambiguously increase everywhere if amenities improve in one city. In the Cobb-Douglas case, $\log \overline{v}$ increases proportionally to the average of $\log \theta$ across cities, so that λ_θ coincides with the welfare impact of a one unit improvement in shape in all cities. The assumption of a fixed total population is extreme, as many migrants into cities are coming from the countryside rather than reallocating across cities. The alternative extreme assumption is that of a perfectly elastic supply of migrants to cities, with indirect utility being pinned down by a reservation utility in the countryside, which delivers the prediction that any improvement in amenities will result in larger urban populations but no welfare change. This provides a lower bound of zero for the welfare effect.

In a richer model with heterogeneous households, there will be welfare impacts on inframarginal households, as the Rosen-Roback conditions continue to hold for the marginal household. As discussed

in Moretti (2011), the welfare impacts on inframarginal agents will depend on the relative elasticity of labor and of housing supply. A lower local elasticity of labor implies a larger incidence on households, with the elasticity of local labor supply being ultimately governed by households' idiosyncratic preferences for locations.

With its assumption of homogeneous agents, the model also rules out sorting. The estimated compensating differentials should be thus thought of as an underestimate of true equalizing differences for those with a strong preference for compact layouts, and an overestimate for those with weak preferences.

In the model, any cost and benefit of city shape that requires cross-city compensating differentials will be part of the θ and A bundles. However, the model is agnostic on the specific channels. City shape may reduce utility because of worse service delivery in disconnected cities or because of worse accessibility. The latter includes direct costs of commuting but also "indirect" costs stemming from traffic congestion, other externalities (such as the disutility from traffic noise or pollution), or other utility costs borne to cope with bad shape - for example, households may give up certain trips or may choose to live, work, or shop in less preferred locations in order to avoid lengthy commutes. Modelling these channels is challenging and distinguishing these different costs in the data would require much more granular data at the sub-city level than what is available for India.

The costs of bad shape that are directly related to accessibility could be accounted for more explicitly by nesting a within-city model in the cross-city framework. In such a framework, some of the costs associated with longer distances could be offset at the sub-city level by the local rent gradient. In Section C in this Appendix I provide a sketch of such a model. In a monocentric, open city with topographic constraints, households directly pay for commuting costs (that are linear in distance) out of their budgets. The predictions for rents and population are consistent with my empirical findings.

The housing sector in the model features constant housing supply elasticity ($1/(\delta - 1)$) across cities. Allowing for heterogeneity in housing supply elasticity across cities would not change the predicted sign of the relationship between θ, A, and the endogenous variables. However, the magnitudes would be affected: to the extent that good shape affects households' indirect utility in equilibrium, in more inelastic cities the impacts on population and wages would be attenuated and the impact on rents would be amplified (in absolute terms). The current framework could be extended allowing housing supply elasticity to be jointly determined with city shape via geography, echoing the insights of Saiz (2010). This would provide a richer characterization of the relationship between shape and supply elasticity, but empirically disentangling the supply elasticity impact would require additional orthogonal sources of variation.

While the notion of compensating differentials based on rents and wages and the model's reduced-form predictions are very general, the calculation of λ_θ and λ_A rely on particular functional forms. This framework uses standard Cobb-Douglas functional form assumptions, which imply homothetic preferences and a constant housing expenditure share. The latter assumption is in line with a large literature (e.g. Ahlfeldt et al. (2015)) and finds empirical support in the U.S. (Davis and Ortalo-Magné 2011). Whether this is plausible for developing countries is less clear and could be addressed in future

work. *A priori*, it is also unclear whether the marginal willingness to pay for "good shape" should be the same across income levels. Poorer households without access to individual means of transportation may be the ones affected the most by bad shape, in line with my findings on slum dwellers making a larger share of the population in compact cities.

Finally, the model does not allow for externalities. In the presence of congestion in consumption, $\widehat{\lambda_\theta}$ would be capturing the equilibrium effect of shape, gross of congestion, providing a lower bound for λ_θ. Similarly, in the presence of agglomeration in production, production amenities will affect productivity both directly, through λ_A, and indirectly, through their effect on city size N. If compact cities have larger populations, this will make them more productive through agglomeration, which will amplify the direct productivity impact of compactness. In this case, estimates of the productivity impact of shape will be an upper bound for λ_A.

C. Spatial Equilibrium within the City and Topographic Constraints

In this Section I provide a framework that embeds irregular city shape in a model of spatial equilibrium within, rather than across cities. I present a simple version of a monocentric city model, augmented with topographic obstacles. This allows me to focus on the implications of city shape for the distribution of households within a city and for commuting. While data availability constraints prevent me from taking this model to the data, the cross-city implications of this model are consistent with my reduced-form results. The within-city model predicts that, for given transportation costs, constrained cities are characterized by a lower population, and by average rents that may be lower or higher depending on the location and the magnitude of the constraint.

Model setup

I draw on a simple version of the monocentric city model (Alonso 1964; Mills 1967; Muth 1969; Brueckner 1987) in which city inhabitants all commute to the CBD. I consider an "open city" version of this model, in which the population of each city is endogenously determined in a way that ensures spatial equilibrium across cities. Each individual earns a wage w and consumes L units of land, both of which are fixed across locations. City dwellers face linear commuting costs τd, where d is the distance from the CBD at which they choose to live. The rental cost per unit of land at a distance d from the CBD is $r(d)$, which is endogenously determined in the model. The utility function of city dwellers is $U(C,L)$ where consumption C is equal to wage income net of housing and commuting costs or $W-\tau d-r(d)L$. For a given city choice, inhabitants choose at which distance from the CBD to live by solving the following maximization problem:

$$\max_d U(w-\tau d-r(d)L,L) \tag{C.1}$$

which yields the Alonso-Muth condition as the first-order condition:

$$r'(d)=-\frac{\tau}{L}, \tag{C.2}$$

The rent function in the city is thus

$$r(d)=r(0)-\frac{\tau}{L}d \tag{C.3}$$

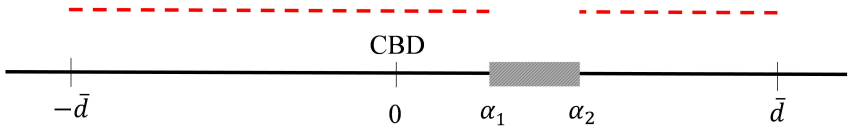


Figure C.1: Population in a linear city with a constraint

with rents declining with distance in a way that offsets the increase in transportation costs. Utility is equalized at any distance from the CBD.

Equilibrium in the constrained city

Now let us make some assumptions on the geometry of the city, and consider the case of a city with topographic or planning constraints. For the sake of simplicity and to provide closed-form solutions, I consider a linear city, in which people live on one dimension along a line. The intuitions and qualitative predictions of the model carry over to a two-dimensional city (as in the standard Alonso-Mills-Muth framework).

In the benchmark model without topographic constraints, individuals can locate on any point along the line; as a result, the distance-minimizing city structure is one in which inhabitants are symmetrically distributed along the line on either side of the CBD. In contrast, a constrained city is one in which certain locations are undevelopable. I model this by assuming that, on one side of the CBD, locations at distances between α_1 and α_2 from the CBD are unavailable, with $0 < \alpha_1 < \alpha_2$.[23] This layout is illustrated in Figure C.1. The plane in which the city is located is represented as the solid black line. Locations along the line are expressed as distances from the CBD, the position of which is normalized at 0. The constraint is represented by the hatched rectangle. For a given city population, the distance-minimizing city structure in the constrained city may become asymmetric, with a smaller fraction of the population locating on the constrained side of the line. The distribution of inhabitants under this city structure is depicted as the dashed red line in Figure C.1. The edge of the city on either side of the CBD is placed at some distance $\bar{d}$, that will be endogenously determined in the model.[24]

Below, I solve for the equilibrium population and rents in the constrained city. The first step is to solve the model for a city population of size N, which will be then endogenized. Assuming that N is sufficiently large relative to the size of the topographic obstacle,[25] the population in the constrained city will distribute itself around the CBD as in Figure C.1. On both sides of the CBD, the furthest occupied location will be at distance $\bar{d}$. The constrained side of the line, however, offers only $\bar{d} - (\alpha_2 - \alpha_1)$ units of inhabitable land. N residents using L units of land each will require NL units of land in total, that are distributed across the two sides of the CBD.

[23] The model's intuitions also apply to a city with multiple constraints.

[24] The benchmark, unconstrained city can be viewed as a special case of the constrained city for which $\alpha_1 = \alpha_2$ (i.e. the obstacle has size 0).

[25] Specifically, NL has to be greater than $(\alpha_1 + \alpha_2)$, otherwise the constraint will never be reached. This condition will be met provided that the city pays a high enough wage relative to transportation costs.

We thus have $NL = \bar{d} + \bar{d} - (\alpha_2 - \alpha_1)$, implying:

$$\overline{d} = \frac{NL + (\alpha_2 - \alpha_1)}{2}. \tag{C.4}$$

In contrast, in the unconstrained city we would have $NL = 2\overline{d}$, as residents distribute themselves symmetrically on each side of the CBD, all the way to the edge of the city at distance $\overline{d}$.

Next, consider the rent function $r(d)$. Assume that rents at the city edge $\bar{d}$ are equal to $\underline{r}$, the opportunity cost of land. By setting $r(\bar{d}) = \underline{r}$ in (C.3) one can obtain $r(0) = \underline{r} + \frac{\tau}{L}\bar{d}$, which implies:

$$r(d) = \underline{r} + \frac{\tau}{L}\bar{d} - \frac{\tau}{L}d. \tag{C.5}$$

Plugging (C.4) in (C.5) yields:

$$r(d) = \underline{r} + \frac{\tau N}{2} + \frac{\tau(\alpha_2 - \alpha_1)}{2L} - \frac{\tau}{L}d. \tag{C.6}$$

In the open-city framework, N is determined by utility-equalizing population flows across cities. Denoting the reservation utility as $\underline{U}$, spatial equilibrium across cities implies $U(w - \tau d - r(d)L, L) = \underline{U}$. Plugging (C.6) into the utility function, this condition becomes:

$$U\left(w - \underline{r}L - \frac{\tau NL}{2} - \frac{\tau(\alpha_2 - \alpha_1)}{2}, L\right) = \underline{U}. \tag{C.7}$$

Let us now consider average rents in the constrained city. In order to derive simple closed-form solutions for N and $r(d)$, further assume that income net of commuting and housing costs in the reservation location is equal to $\underline{C}$:

$$w - \underline{r}L - \frac{\tau NL}{2} - \frac{\tau(\alpha_2 - \alpha_1)}{2} = \underline{C}. \tag{C.8}$$

From (C.8) one can pin down the equilibrium N:

$$N = \frac{2(w - \underline{r}L - \underline{C})}{\tau L} - \frac{(\alpha_2 - \alpha_1)}{L}. \tag{C.9}$$

Plugging (C.9) into (C.4) yields:

$$\overline{d} = \frac{(w - \underline{r}L - \underline{C})}{\tau} \tag{C.10}$$

which does not depend on the size or on the position of the topographic obstacles. Plugging the equilibrium $\overline{d}$ into (C.5) yields:

$$r(d) = \underline{r} + \frac{w - \underline{r}L - \underline{C}}{L} - \frac{\tau}{L}d. \tag{C.11}$$

Comparative statics

Below I discuss the model's prediction for the equilibrium population and rents in a constrained versus unconstrained city. I show that population is unambiguously lower in constrained cities - a prediction which is borne in my data. On the other hand, whether rents are higher or lower in constrained cities depends on the position of the constraint, making it ultimately an empirical question.

From (C.7) and (C.9) it is apparent that the city's population N is smaller, the larger the size of the constraint $(\alpha_2 - \alpha_1)$. Intuitively, a city with topographic constraints is one in which, for a given maximum distance from the CBD, there are fewer locations available, and in equilibrium it will host a

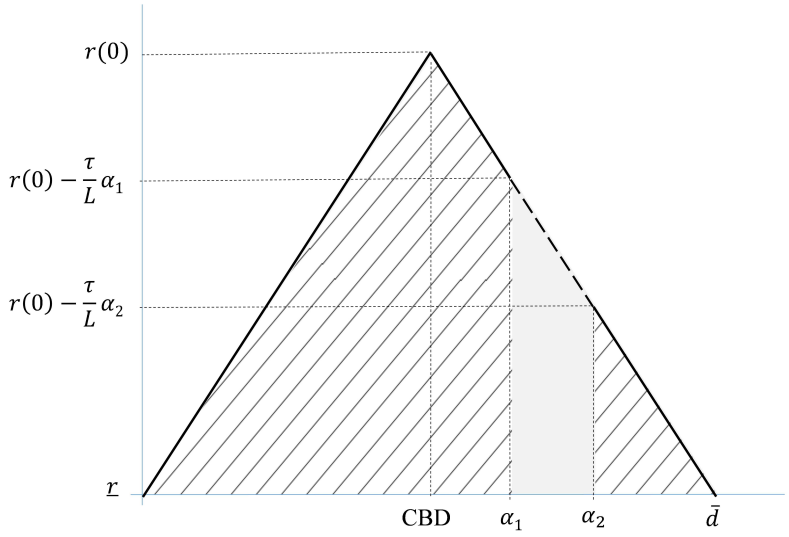


Figure C.2: Rents distribution in a linear city with a constraint

smaller population.

I next show that, all else being equal, average rents in the constrained city may be lower or higher than in the unconstrained city. Consider two cities that are identical in all parameters of the model, except for the fact that one is constrained and the other is unconstrained. The distribution of rents as a function of distance from the CBD in the constrained city is represented by the solid line in Figure C.2. The solid line plus the dashed line segment, taken together, represent rents in the unconstrained city.

Note that both cities have the same rent gradient $r(d)$ and the same equilibrium $\bar{d}$, but the constrained city is missing a portion of the distribution of rents, corresponding to the dashed segment. The hatched area in Figure C.2 corresponds to total rents in the constrained city; the hatched plus the solid area correspond to total rents in the unconstrained city. Rents per unit of land will be higher or lower in the constrained city depending on the size and position of the constraint. Intuitively, if the topographic obstacle precludes development close to the CBD, where rents would be high, average rents will be lower than in the unconstrained city. If the topographic obstacle precludes development far from the CBD, where rents would be low, average rents will be higher than in an unconstrained city. This intuition applies also to cases with multiple constraints that may introduce gaps in the rent distribution at different points.

This can be shown algebraically by computing average rents in the two cities, which can be easily done by calculating the areas of the relevant triangles and rectangles in Figure C.2. Total rents in the unconstrained city, denoted as R_U, can be calculated as:

$$R_U = \frac{2\bar{d}\left[r(0) - \underline{r}\right]}{2} = \frac{\left(w - \underline{r}L - \underline{C}\right)^2}{\tau L}. \tag{C.12}$$

Denoting the equilibrium population in the unconstrained city with N_U, the average rent per unit of occupied land in the unconstrained city is:

$$\frac{R_U}{LN_U} = \frac{R_U \tau}{2\left(w - \underline{r}L - \underline{C}\right)} = \frac{w - \underline{r}L - \underline{C}}{2L}. \tag{C.13}$$

Total rents in the constrained city, denoted as R_C, are equal to total rents in the unconstrained city minus the area of the solid trapezoid in Figure C.2, which we can denote as A:

$$R_C = R_U - A = R_U - (\alpha_2 - \alpha_1)\left[\left(r(0) - \frac{\tau}{L}\alpha_2 - \underline{r}\right) + \frac{\frac{\tau}{L}(\alpha_2 - \alpha_1)}{2}\right]. \tag{C.14}$$

Denoting the equilibrium population in the constrained city with N_C, the average rent per unit of land in the constrained city is thus:

$$\frac{R_C}{LN_C} = \frac{R_U - A}{L\left(N_U - \frac{(\alpha_2 - \alpha_1)}{L}\right)}. \tag{C.15}$$

Average rents in the constrained city are lower than in the unconstrained city when (C.15) is smaller than (C.13), or equivalently when:

$$\alpha_2 - \alpha_1 < \frac{LN_U}{R_U} A. \tag{C.16}$$

Plugging in the expressions for N_U, R_U and A, this inequality simplifies to:

$$\alpha_1 + \alpha_2 < \frac{w - \underline{r}L - \underline{C}}{\tau}. \tag{C.17}$$

Whether a constrained city has lower average rents than an unconstrained city depends on the size of the constraint $(\alpha_2 - \alpha_1)$ and the position of the constraint α_1. All else being equal, when the obstacle is close to the CBD (α_1 is small), the condition above is more likely to be satisfied; intuitively, topography is preventing development in a location that would be a high-rent one due to its proximity to the center. Furthermore, for a given topography, the condition above is more likely to be satisfied when wages are higher or transportation costs are lower. High wages and low transportation costs attract a larger population and make the city more spread out (leading to a larger $\bar{d}$); as a result, locations at distance α_1 from the CBD become relatively more central and demand a higher rent.

Note that the analysis above holds transportation costs constant across constrained and unconstrained cities. In a richer model, one could assume that transportation costs per unit distance are higher in cities that have irregular layouts due to topographic constraints. A standard comparative statics result in the open-city version of the monocentric city model is that cities with higher transportation costs have lower rents and smaller populations (Brueckner 1987), which would further align the theoretical predictions with my empirical findings.

D. Signing the OLS Bias

In this Section I follow up on the discussion in Section 5 in the paper and illustrate analytically that the OLS bias from estimating the impact of city shape on population and other outcomes has an ambiguous sign. I also discuss under what conditions the bias for the impact of shape on population is positive.

Denoting city population with N and city shape with S, consider the following version of the

estimating equation, where for simplicity I have dropped subscripts and additional regressors:

$$logN = \alpha_1 S + u_1 \tag{D.1}$$

In what follows, I show analytically that $cov(S, u_1) \neq 0$. City shape is the result of an interaction of exogenous determinants (such as topographic obstacles) and endogenous determinants. While the endogenous determinants are manifold, to fix ideas, consider two factors: local institutional capacity and highways connecting cities. With good local institutional capacity, urban planners can encourage compact development through well-enforced master plans and land use regulations. Highways connecting into a city can affect urban form by encouraging urban development along transit corridors, which has been associated with sprawl (Baum-Snow 2007) and potentially deteriorates urban shape. In Section 5 in the paper I discuss more factors that affect city shape and that would generate selection effects similar to those highlighted here. Denoting (exogenous) geographic predictors of city shape as $\tilde{S}$, local institutional capacity as $Inst$, and road infrastructure as $Infra$, S can be written as a function of its determinants:

$$S = \beta \tilde{S} + \delta_1 Inst + \delta_2 Infra + \eta \tag{D.2}$$

Assume that $\beta > 0$ (potential shape predicts actual shape), $\delta_1 < 0$ (better institutional capacity makes cities more compact), and $\delta_2 > 0$ (highways cause sprawl). Further assume that $cov(Inst, \eta) = cov(Infra, \eta) = cov(S, \eta) = 0$.

The endogeneity problem associated with $Inst$ and $Infra$ stems from the fact that city size N affects both local institutional capacity and highways. Larger cities have both greater institutional capacity and more infrastructural investment. We thus have:

$$Inst = \gamma_1 logN + \xi_1 \tag{D.3}$$

$$Infra = \gamma_2 logN + \xi_2 \tag{D.4}$$

where $\gamma_1 > 0$ and $\gamma_2 > 0$ and $cov(logN, \xi_1) = cov(logN, \xi_2) = 0$.

Plugging (D.3) and (D.4) into (D.2) one obtains:

$$S = \beta \tilde{S} + \alpha_2 logN + \tilde{\eta} \tag{D.5}$$

where $\alpha_2 = \delta_1 \gamma_1 + \delta_2 \gamma_2$ and $\tilde{\eta} = \delta_1 \xi_1 + \delta_2 \xi_2 + \eta$.

In equilibrium, city shape is a function of exogenous geographic predictors plus a term that depends endogenously on population. The sign of α_2 is ambiguous because $\delta_1 < 0$ and $\delta_2 > 0$. If the effect of highways is stronger than that of institutional capacity, α_2 will be positive. The descriptive patterns in Table A2 as well as the OLS impacts in Table3 indicate a positive correlation between city size N and shape S, suggesting that empirically α_2 is positive.

Solving the system of equations given by (D.5) and (D.1) one obtains:

$$S = \frac{\beta}{1 - \alpha_2 \alpha_1} \tilde{S} + \frac{\alpha_2 u_1}{1 - \alpha_2 \alpha_1} + \frac{\tilde{\eta}}{1 - \alpha_2 \alpha_1}. \tag{D.6}$$

Going back to estimating equation (D.1), one can now compute the covariance between the error

term u_1 and S as follows:

$$cov(S,u_1) = cov\Big(\frac{\alpha_2}{1-\alpha_2\alpha_1}u_1, u_1\Big) = \sigma^2_{u_1}(\frac{\alpha_2}{1-\alpha_2\alpha_1}). \tag{D.7}$$

In general, (D.7) has an ambiguous sign. As a result, the OLS bias in estimating the impact of S in (D.1) could be positive or negative. However, if $\alpha_2 > 0$ (the equilibrium correlation between shape and population is positive) and $\alpha_1 < 0$ (the structural effect of shape on population is negative), then $(\frac{\alpha_2}{1-\alpha_2\alpha_1})$ will be unambiguously positive and the OLS estimate will be biased towards positive values. This is indeed what the estimates show, with negative IV impacts and positive OLS impacts of city shape on population growth.

E. Single-Instrument Approach

In this Section I outline an alternative implementation of my instrumental variables strategy, which I employ in the robustness check presented in Table 8 and discussed in Section 7.

Recall from Section 5 in the paper that in the benchmark estimation I consider city shape and area as two endogenous regressors, and I employ as instrumental variables "potential" city shape and projected historical population. Potential shape is calculated based on the relative position of topographic obstacles encountered as a city grows along a predicted expansion path, which in turn depends on the city's own projected population growth. Below I present an alternative strategy that does not rely on projected historical population. Relative to the baseline approach, this entails two differences: first, the instrument is constructed using a completely mechanical model for city expansion that postulates the same rate of expansion for all cities. Second, city area is not directly controlled for in the estimating equations, but instead both right- and left-hand side variables are normalized by area.

The first step is to determine $\widehat{r}_{c,t}$, i.e. the predicted city radius within which the potential footprint is constructed. Under this alternative approach I do so by postulating that cities expand at the same rate, equivalent to the average expansion rate across all cities in the sample. Specifically, the steps involved are the following:

(i) Denoting the area of city c's actual footprint in year t as $area_{c,t}$, I pool together the 1951-2010 panel of cities and estimate the following regression:

$$\log(area_{c,t}) = \theta_c + \gamma_t + \varepsilon_{c,t} \tag{E.1}$$

where θ_c and γ_t denote city and year fixed effects.

(ii) From the regression above, I obtain $\widehat{area_{c,t}}$, and corresponding $\widehat{r_{c,t}} = \sqrt{\frac{\widehat{area_{c,t}}}{\pi}}$.

The second modification relative to the baseline approach is in the estimating equations. Instead of controlling for city area explicitly, this approach relies on normalizing both right- and left-hand side variables by city area, regressing population density on the normalized version of the shape index.

Define population density[26] as

$$d_{c,t} = \frac{pop_{c,t}}{area_{c,t}}$$

[26]Note that this does not coincide with population density as defined by the Census, which reflects administrative boundaries.

and denote the normalized version of shape as nS.

The alternative estimating equation becomes:

$$d_{c,t} = a \cdot nS_{c,t} + \mu_c + \rho_t + \eta_{c,t} \tag{E.2}$$

which includes the endogenous regressor $nS_{c,t}$. This is a counterpart of equation (8) in the paper.

The corresponding first-stage equation is

$$nS_{c,t} = \beta \cdot \widetilde{nS_{c,t}} + \lambda_c + \gamma_t + \varepsilon_{c,t}. \tag{E.3}$$

where $\widetilde{nS_{c,t}}$ is namely the normalized shape index computed for the potential footprint.

F. Nonparametric Employment Subcenter Identification (McMillen 2001)

In order to compute the number of employment subcenters in each city, used as the dependent variable in Table 10, I employ the two-stage, non-parametric approach described in McMillen (2001). This procedure identifies employment subcenters as locations that have significantly larger employment density than nearby ones, and that have a significant impact on the overall employment density function in a city. The data on firms' location used as input in this procedure is discussed in Section A6 above.

The procedure outlined below is performed separately for each city in the 2005 sample. As units of observation within each city, I consider grid cells of 0.01 degree latitude by 0.01 degree longitude, with an area of approximately one square kilometer. I calculate a proxy for employment density in each cell, by considering establishments from the 2005 Economic Census located in that cell and summing their reported number of employees.[27] In order to define the CBD using a uniform criterion for all cities, I consider the centroid of the 1951 footprint. Results are similar using the 2005 centroid as an alternative definition.

In the first stage of this procedure, "candidate" subcenters are identified as those grid cells with significant positive residuals in a smoothed employment density function. Let y_i be the log employment density in grid cell i; denote its distance from the CBD with x and the error term with ε_i, I estimate:

$$y_i = f(x) + \varepsilon_i \tag{F.1}$$

using locally weighted regression, employing a tricube kernel and a 50% window size. This flexible specification allows for local variations in the density gradient, which are likely to occur in cities with topographic obstacles. Denoting with $\widehat{y_i}$ the estimate of y for cell i, and with $\widehat{\sigma_i}$ the corresponding standard error, candidate subcenters are grid cells such that $(y_i - \widehat{y_i}) / \widehat{\sigma_i} > 1.96$.

The second stage of the procedure selects those locations, among candidate subcenters, that have significant explanatory power in a semiparametric employment density function estimation. Let D_{ij} be the distance between cell i and candidate subcenter j, and denote with $DCBD_i$ the distance between

[27] The Directory of Establishments provides establishment-level employment only by broad categories, indicating whether the number of employees falls in the 10-50, 51-100, or 101-500 range, or is larger than 500. In order to assign an employment figure to each establishment, I consider the lower bound of the category.

cell i and the CBD. With S candidate subcenters, denoting the error term with u_i, the semi-parametric regression takes the following form:

$$y_i = g(DCBD_i) + \sum_{j=1}^{S} \delta_j^1 (D_{ji})^{-1} + \delta_j^2 (-D_{ji}) + u_i. \tag{F.2}$$

In the specification above, employment density depends non-parametrically on the distance to the CBD, and parametrically on subcenter proximity, measured both in levels and in inverse form. This parametric specification allows us to conduct convenient hypothesis tests on the coefficients of interest δ_j^1 and δ_j^2. (F.2) is estimated omitting cells i corresponding to one of the candidate subcenters or to the CBD. I approximate $g(.)$ using cubic splines.

If j is indeed an employment subcenter, the variables $(D_j)^{-1}$and/or $(-D_j)$ should have a positive and statistically significant impact on employment density y. One concern with estimating (F.2) is that, with a large number of candidate subcenters, the distance variables D_{ij} can be highly multicollinear. To cope with this problem, a stepwise procedure is used to select which subcenter distance variables to include in the regression. In the first step, all distance variables are included. At each step, the variable corresponding to the lowest t-statistic is dropped from the regression, and the process is repeated until all subcenter distance variables in the regression have a positive coefficient, significant at the 20% level. The final list of subcenters includes the sites with positive coefficients on either $(D_j)^{-1}$ or $(-D_j)$.

References

[1] Adhvaryu, Achyuta, Amalavoyal Chari, and Siddharth Sharma. 2013. "Replication data for: Firing Costs and Flexibility: Evidence from Firms' Employment Responses to Shocks in India." *The MIT Press* [publisher], Harvard Dataverse [distributor]. https://doi.org/10.7910/DVN/24587.

[2] Ahlfeldt, Gabriel M., Stephen J. Redding, Daniel M. Sturm, and Nikolaus Wolf. 2015. "The Economics of Density: Evidence From the Berlin Wall." *Econometrica* 83(6): 2127-89. https://doi.org/10.3982/ECTA10876.

[3] Akbar, Prottoy, Victor Couture, Gilles Duranton, and Adam Storeygard. 2018. "Accessibility in Urban India." Unpublished.

[4] Akbar, Prottoy, Victor Couture, Gilles Duranton, and Adam Storeygard. 2019. "Mobility and Congestion in Urban India." NBER Working Papers 25218.

[5] Alonso, William. 1964. *Location and Land Use*. Cambridge: Harvard University Press.

[6] Angel, Shlomo, Daniel L. Civco, and Jason Parent. 2009. "Shape Metrics." https://clear.uconn.edu/tools/Shape_Metrics/ESRI_2009_Shape%20Metrics.ppt (accessed March 19, 2020).

[7] Angel, Shlomo, Daniel L. Civco, and Jason Parent. 2010. "Ten Compactness Properties of Circles: Measuring Shape in Geography." *The Canadian Geographer/Le Géographe Canadien* 54(4): 441-61. https://doi.org/10.1111/j.1541-0064.2009.00304.x.

[8] Balk, Deborah, Uwe Deichmann, Greg Yetman, Francesca Pozzi, Simon Hay, and Andy Nelson. 2006. "Determining Global Population Distribution: Methods, Applications and Data." *Advances in Parasitology* 62: 119-156. https://doi.org/10.1016/S0065-308X(05)62004-0

[9] Barr, Jason, Troy Tassier, and Rossen Trendafilov. 2011. "Depth to Bedrock and the Formation of the Manhattan Skyline, 1890-1915." *The Journal of Economic History* 71(4): 1060-77. https://doi.org/10.1017/S0022050711002245.

[10] Baruah, Neeraj, John Vernon Henderson, and Cong Peng. 2017. "Colonial Legacies: Shaping African Cities." SERC Urban and Spatial Programme Discussion Paper 0226.

[11] Baum-Snow, Nathaniel. 2007. "Did Highways Cause Suburbanization?" *The Quarterly Journal of Economics* 122(2): 775-805. https://doi.org/10.1162/qjec.122.2.775.

[12] Bertaud, Alain, and Jan K. Brueckner. 2005. "Analyzing Building-Height Restrictions: Predicted Impacts and Welfare Costs." *Regional Science and Urban Economics* 35(2): 109-25. https://doi.org/10.1016/j.regsciurbeco.2004.02.004.

[13] Brueckner, Jan K. 1987. "The Structure of Urban Equilibria: A Unified Treatment of the Muth-Mills Model." In *The Handbook of Regional and Urban Economics*, Volume 2, edited by Edwin S. Mills. Amsterdam: North Holland Press.

[14] Carroll, Mark, Charlene DiMiceli, Margaret Wooten, Alfred Hubbard, Robert Sohlberg, John Townshend. 2009. "MOD44W MODIS/Terra Land Water Mask Derived from MODIS and SRTM L3 Global 250m SIN Grid V005." NASA EOSDIS Land Processes DAAC [distributor]. https://www.tandfonline.com/doi/full/10.1080/17538940902951401.

[15] Central Statistics Office. 1990-1991. "Annual Survey of Industries (India), Summary." Ministry of Statistics and Programme Implementation. Provided by CSO(IS Wing) Kolkata.

[16] Central Statistics Office. 2009-2010. "Annual Survey of Industries (India), Unit Level Data." Ministry of Statistics and Programme Implementation. Provided by CSO(IS Wing) Kolkata.

[17] Centre for Industrial and Economic Research (CIER). 1990. "Industrial Databook 1990." New Delhi: CIER.

[18] Center for International Earth Science Information Network, CUNY Institute for Demographic Research, International Food Policy Research Institute, The World Bank, and Centro Internacional de Agricultura Tropical. 2017. "Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Settlement Points, Revision 01." Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4BC3WG1 (accessed February 27, 2020).

[19] Chauvin, Juan Pablo, Edward Glaeser, Yueran Ma, and Kristina Tobio. 2017. "What Is Different about Urbanization in Rich and Poor Countries? Cities in Brazil, China, India and the United States." *Journal of Urban Economics* 98: 17-49. https://doi.org/10.1016/j.jue.2016.05.003.

[20] Chen, Xi, and William D. Nordhaus. 2016. "Global Gridded Geographically Based Economic Data (G-Econ), Version 4." Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). Retrieved from: https://gecon.yale.edu/data-and-documentation-g-econ-project (accessed March 13, 2020).

[21] Davis, Morris A., and François Ortalo-Magné. 2011. "Household Expenditures, Wages, Rents." *Review of Economic Dynamics* 14(2): 248-61. https://doi.org/10.1016/j.red.2009.12.003.

[22] Desai, Sonalde, Reeve Vanneman, and National Council of Applied Economic Research, New Delhi. 2005. "India Human Development Survey (IHDS)." Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR22626.v12

[23] Desai, Sonalde, Reeve Vanneman, and National Council of Applied Economic Research, New Delhi. 2012. "India Human Development Survey-II (IHDS-II)." Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor]. https://doi.org/10.3886/ICPSR36151.v6

[24] Dev, Satvik. 2006. "Rent Control Laws in India: A Critical Analysis." Centre for Civil Society Working Paper No. 158.

[25] FAO/IIASA. 2010. "Global Agro-ecological Zones (GAEZ v3.0)." FAO, Rome, Italy and IIASA, Laxenburg, Austria. Retrieved from: http://gaez.fao.org/Main.html# (accessed March 4, 2020).

[26] Fernandes, Ana, and Gunjan Sharma. 2012. "Determinants of Clusters in Indian Manufacturing: The Role of Infrastructure, Governance, Education, and Industrial Policy." IGC working paper C-35005-INC-1.

[27] Glaeser, Edward. 2008. *Cities, Agglomeration and Spatial Equilibrium*. Oxford: Oxford University Press.

[28] Greenstone, Michael, and Rema Hanna. 2014. "Environmental Regulations, Air and Water Pollution, and Infant Mortality in India." *American Economic Review* 104(10): 3038-72. https://doi.org/10.1257/aer.104.10.3038.

[29] Henderson, Mark, Emily T. Yeh, Peng Gong, Christopher Elvidge, and Kim Baugh. 2003. "Validation of Urban Boundaries Derived from Global Night-Time Satellite Imagery." *International Journal of Remote Sensing* 24(3): 595-609. https://doi.org/10.1080/01431160304982.

[30] Hengl T, de Jesus JM, MacMillan RA, Batjes NH, Heuvelink GBM, Ribeiro E, et al. 2014. "Data from: SoilGrids1km - Global Soil Information Based on Automated Mapping." *PLoS ONE* 9(8): e105992. Retrieved from: http://www.soilgrids.org/ (accessed June 2018).

[31] International Monetary Fund. 1970-. "World Economic Outlook." Washington, DC: International Monetary Fund (IMF). Retrieved from: http://www.econstats.com/weo/CIND.htm (accessed March 2, 2020).

[32] Iyer, Lakshmi. 2010. "Direct versus Indirect Colonial Rule in India: Long-Term Consequences." *Review of Economics and Statistics* 92(4): 693-713. https://doi.org/10.1162/REST_a_00023.

[33] Joshi, Pawan K., Brij M. Bairwa, Richa Sharma, Vinay Sinha. 2011. "Assessing Urbanization Patterns over India Using Temporal DMSP-OLS Night-time Satellite Data." *Current Science* 100(10), 1479-1482. https://www.currentscience.ac.in/Volumes/100/10/1479.pdf (accessed March 18, 2019).

[34] Kumar, Hemanshu, and Rohini Somanathan. 2009. "Mapping Indian Districts Across Census Years, 1971-2001." *Economic and Political Weekly* 44 (41/42): 69-73. www.jstor.org/stable/25663682.

[35] McMillen, Daniel P. 2001. "Nonparametric Employment Subcenter Identification." *Journal of Urban Economics* 50(3): 448-73. https://doi.org/10.1006/juec.2001.2228.

[36] McMillen, Daniel P., and Stefani C. Smith. 2003. "The Number of Subcenters in Large Urban Areas." *Journal of Urban Economics* 53 (3): 321-38. https://doi.org/10.1016/S0094-1190(03)00026-3.

[37] Mills, Edwin S. 1967. "An Aggregative Model of Resource Allocation in a Metropolitan Area." *American Economic Review* 57(2), 197-210. www.jstor.org/stable/1821621.

[38] Ministry of Urban Affairs. 1981-2001. "Identified/Estimated Slum Population." Government of India. Retrieved from: https://www.indiastat.com/ (accessed February 27, 2020).

[39] Mitra, Ashok. 1980. *Population and Area of Cities, Towns, and Urban Agglomerations, 1872-1971*. Bombay: Allied Publishers.

[40] Moretti, Enrico. 2011. "Local Labor Markets." In *Handbook of Labor Economics*, Volume 4, Part B, edited by David Card, and Orley Ashenfelter. Amsterdam: North-Holland.

[41] Munshi, Kaivan, and Mark Rosenzweig. 2016. "Networks and Misallocation: Insurance, Migration, and the Rural-Urban Wage Gap." *American Economic Review* 106(1): 46-98. https://doi.org/10.1257/aer.20131365.

[42] Muth, Richard F. 1969. *Cities and Housing*. Chicago: University of Chicago Press.

[43] National Aeronautics and Space Administration (NASA), and Japan's Ministry of Economy, Trade, and Industry (METI). 2011. "Aster Global Digital Elevation Model, v002." NASA EOSDIS Land Processes DAAC [distributor]. https://doi.org/10.5067/ASTER/ASTGTM.002 (accessed March 4, 2020).

[44] National Geophysical Data Center. 1992-. "DMSP-OLS Nighttime Lights." Silver Spring, MD: National Oceanic and Atmospheric Administration. Retrieved from: https://ngdc.noaa.gov/eog/dmsp/downloadV4composites.html (accessed March 4, 2020).

[45] National Sample Survey Office. 2005-2006. "National Sample Survey (NSS) on 62nd Round, Sch. 1.0: (consumer Expenditure) (unit level)." New Delhi: Government of India, Ministry of Statistics and Programme Implementation.

[46] National Sample Survey Office. 2007-2008. "National Sample Survey (NSS) on 64th Round, Sch. 1.0: (consumer Expenditure) (unit level)." New Delhi: Government of India, Ministry of Statistics and Programme Implementation.

[47] Office of the Registrar General & Census Commissioner, India. 1981. "Census of India: Town directory, part X-A (i)." New Delhi: Ministry of Home Affairs, Government of India.

[48] Office of the Registrar General & Census Commissioner, India. 1981-2011. "Population Totals by Urban Agglomeration." New Delhi: Ministry of Home Affairs, Government of India. Retrieved from Thomas Brinkhoff: City Population; http://www.citypopulation.de/India.html (accessed June 2014).

[49] Office of the Registrar General & Census Commissioner, India. 1991a. "Census of India 1991: Primary Census Abstract, Town Directory." New Delhi: Ministry of Home Affairs, Government of India.

[50] Office of the Registrar General & Census Commissioner, India. 1991b. "Census of India 1991: Primary Census Abstract, H series tables." New Delhi: Ministry of Home Affairs, Government of India.

[51] Office of the Registrar General & Census Commissioner, India. 2001a. "Census of India, 2001: Primary Census Abstract." New Delhi: Ministry of Home Affairs, Government of India.

[52] Office of the Registrar General & Census Commissioner, India. 2001b. "Census of India 2001: H Series Tables." New Delhi: Ministry of Home Affairs, Government of India.

[53] Office of the Registrar General & Census Commissioner, India. 2001c. "Census of India 2001: Code List for Land Regions (State, District, Sub-District, Town, Village)." New Delhi: Ministry of Home Affairs, Government of India. https://censusindia.gov.in/Census_Data_2001/PLCN/plcn.html (accessed March 19, 2020).

[54] Office of the Registrar General & Census Commissioner, India. 2001d. "Census of India 2001: village level amenities data." New Delhi: Ministry of Home Affairs, Government of India.

[55] Office of the Registrar General & Census Commissioner, India. 2001e. "Census of India 2001: village and town location codes." New Delhi: Ministry of Home Affairs, Government of India.

[56] Office of the Registrar General & Census Commissioner, India. 2001f. "Table - 6: Population, Population in the age group 0-6 and literates by sex - Urban Agglomeration/Town: 2001" New Delhi: Ministry of Home Affairs, Government of India. http://web.archive.org/web/20040615023133/; http://www.censusindia.net/results/ua_town.php?stad=A&balls=allballs (accessed March 19, 2020).

[57] Office of the Registrar General & Census Commissioner, India. 2001g. "Table - 3: Population, population in the age group 0-6 and literates by sex - Cities/Towns (in alphabetic order): 2001." New Delhi: Ministry of Home Affairs, Government of India. http://web.archive.org/web/20040616075334/; http://www.censusindia.net/results/town.php?stad=A&state5=999 (accessed March 19, 2020).

[58] Office of the Registrar General & Census Commissioner, India. 2001h. "Census of India 2001: Tables on Houses, Household Amenities and Assets; Table H13: Number of Households Availing Banking Services and Number of Households Having Each of The Specified Asset." New Delhi: Ministry of Home Affairs, Government of India. https://censusindia.gov.in/DigitalLibrary/TablesSeries2001.aspx (accessed March 19, 2020).

[59] Office of the Registrar General, India. 2005. "Economic Census of India 2005: Directories of Establishment." Ministry of Statistics and Programme Implementation, Government of India. Retrieved from: http://www.mospi.gov.in/directories-establishment-fifth-economic-census-2005 (accessed March 05, 2020).

[60] Office of the Registrar General & Census Commissioner, India. 2011a. "Census of India 2011: Primary Census Abstract." New Delhi: Ministry of Home Affairs, Government of India. https://censusindia.gov.in/DigitalLibrary/Tables.aspx (accessed March 19, 2020).

[61] Office of the Registrar General & Census Commissioner, India. 2011b. "Census of India 2011: List of Villages/Towns." New Delhi: Ministry of Home Affairs, Government of India. https://censusindia.gov.in/2011census/Listofvillagesandtowns.aspx (accessed March 19, 2020).

[62] Office of the Registrar General & Census Commissioner, India. 2011c. "Administrative Atlas of India, Census of India 2011." New Delhi: Ministry of Home Affairs, Government of India http://censusindia.gov.in/2011census/maps/administrative_maps/Final%20Atlas%20India%202011.pdf (accessed March 19, 2020).

[63] Office of the Registrar General & Census Commissioner, India. 2011d. "Census of India 2011: Houselisting & Housing Census Tables, Table Hh-14: Percentage Of Households To Total Households By Amenities And Assets." New Delhi: Ministry of Home Affairs, Government of India. http://censusindia.gov.in/2011census/HLO/HL_PCA/Houselisting-housing-HLPCA.html (accessed March 19, 2020).

[64] Office of the Registrar General & Census Commissioner, India. 2011e. "Census of India 2011: B-28: 'Other Workers' By Distance From Residence To Place Of Work And Mode Of Travel To Place Of Work." New Delhi: Ministry of Home Affairs, Government of India. http://censusindia.gov.in/2011census/B-series/B_28.html (accessed March 19, 2020).

[65] Office of the Registrar General & Census Commissioner, India. 2011f. "Census of India 2011: Town amenities tables." New Delhi: Ministry of Home Affairs, Government of India. https://censusindia.gov.in/2011census/dchb/DCHB.html (accessed March 19, 2020).

[66] Office of the Registrar General & Census Commissioner, India. 2011g. "Census of India 2011: Primary Census Abstract Data for Slum (India & States/UTs Town Level) (Excel Format)." New Delhi: Ministry of Home Affairs, Government of India. http://censusindia.gov.in/2011-Documents/slum_data_census_2011.xls (accessed March 19, 2020).

[67] Office of the Registrar General & Census Commissioner, India. 2011h. "Census of India 2011: HH-11 Households Classified by Source and Location of Drinking Water and Availablity of Electricity and Latrine." New Delhi: Ministry of Home Affairs, Government of India. http://censusindia.gov.in/2011census/Hlo-series/HH11.html (accessed March 19, 2020).

[68] OpenStreetMap contributors. 2019. "OpenStreetMap." https://www.openstreetmap.org/ (accessed March 2019).

[69] Patterson, Tom, and Nathaniel V. Kelso. 2012. "Natural Earth v2.0.0." http://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-rivers-lake-centerlines (accessed March 4, 2020).

[70] Roback, Jennifer. 1982. "Wages, Rents, and the Quality of Life." *Journal of Political Economy* 90(6): 1257-78. https://doi.org/10.1086/261120.

[71] Rosen, Sherwin. 1979. "Wage-Based Indexes of Urban Quality of Life." In *Current Issues in Urban Economics*, edited by Peter Mieszkowski. Baltimore: Johns Hopkins University Press.

[72] Roychowdhury, Koel, Simon D. Jones, and Colin Arrowsmith. 2009. "Assessing the Utility of DMSP/OLS Night-time Images for Characterizing Indian Urbanization." Paper presented at the 2009 Joint Urban Remote Sensing Event, Shanghai, China.

[73] Saiz, Albert. 2010. "The Geographic Determinants of Housing Supply." *Quarterly Journal of Economics* 125(3): 1253-96. https://doi.org/10.1162/qjec.2010.125.3.1253.

[74] Small, Christopher, Francesca Pozzi, and Christopher Elvidge. 2005. "Spatial Analysis of Global Urban Extent from DMSP-OLS Night Lights." *Remote Sensing of Environment* 96(3-4): 277-91. https://doi.org/10.1016/j.rse.2005.02.002.

[75] Small, Christopher, and Christopher D. Elvidge. 2013. "Night on Earth: Mapping Decadal Changes of Anthropogenic Night Light in Asia." *International Journal of Applied Earth Observation and Geoinformation* 22: 40-52. https://doi.org/10.1016/j.jag.2012.02.009.

[76] Sridhar, Kala S. 2010. "Impact of Land Use Regulations: Evidence from India's Cities." *Urban Studies* 47(7): 1541-69. https://doi.org/10.1177/0042098009353813.

[77] Transport Research Wing. 1971-2020a. "Urban Road Length for All Indian States." Ministry of Road Transport and Highways, Government of India. Retrieved from CMIE: https://statesofindia.cmie.com/ (accessed February 26, 2020).

[78] Transport Research Wing. 1971-2020b. "Registered Motor Vehicles for All Indian States." Ministry of Road Transport and Highways, Government of India. Retrieved from CMIE: https://statesofindia.cmie.com/ (accessed February 26, 2020).

[79] U.S. Army Map Service. 1955-. "India and Pakistan 1:250,000 Series U502." Courtesy of the University of Texas Libraries, The University of Texas at Austin. https://legacy.lib.utexas.edu/maps/ams/india/ (accessed March 4, 2020).

[80] U.S. Geological Survey. 2005. "Mineral Resources Data System." Reston, Virginia: U.S. Geological Survey. Retrieved from: https://mrdata.usgs.gov/mrds/geo-inventory.php (accessed March 4, 2020).

[81] Wessel, P., and W.H.F. Smith. 2013. "A Global Self-consistent, Hierarchical, High-resolution Geography Database (Version 2.2.2)." SOEST, University of Hawaii, Honolulu, HI and NOAA Geosciences Lab, National Ocean Service, Silver Spring, MD. Retrieved from: ftp://ftp.soest.hawaii.edu/gshhg/legacy/gshhg-shp-2.2.2.zip (accessed March 4, 2020).

[82] World Wildlife Fund and the Center for Environmental Systems Research, University of Kassel. 2004. "Global Lakes and Wetlands Database: Small Lake Polygons (Level 2)." Washington, DC: WWF. Retrieved from: https://www.worldwildlife.org/pages/global-lakes-and-wetlands-database (accessed March 4, 2020).

Table A1: List of most and least compact cities

Rank	City	Shape (normalized)
	Top 10 most compact cities	
1	Rajkot, Gujarat	0.924
2	Kannur, Kerala	0.934
3	Bhopal, Madhya Pradesh	0.943
4	Lucknow, Uttar Pradesh	0.943
5	Meerut, Uttar Pradesh	0.943
6	Thrissur, Kerala	0.945
7	Nashik, Maharashtra	0.948
8	Jaipur, Rajasthan	0.948
9	Jabalpur, Madhya Pradesh	0.952
10	Gwalior, Madhya Pradesh	0.956
	Top 10 least compact cities	
1	Asansol, West Bengal	1.625
2	Jharia-Dhanbad, Jharkhand	1.180
3	Kolkata, West Bengal	1.128
4	Ludhiana, Punjab	1.124
5	Surat, Gujarat	1.111
6	Aurangabad, Maharashtra	1.108
7	Visakhapatnam, Ahndra Pradesh	1.108
8	Patna, Bihar	1.100
9	Amritsar, Punjab	1.100
10	Chennai, Tamil Nadu	1.081

Note: Sample of cities with million-plus population in 2011, ranked by normalized shape in 2010. The normalized shape index has a mean of 0.96 and a standard deviation of 0.07 in 2010.

Table A2: Descriptive correlations

	(1)	(2)	
Dependent variable:	*Shape, 2010*	*Δ Shape, 2010-1950*	*Obs.*
Panel A: 1951 city size quartiles			
Quartile I	-1.188***	-1.152***	351
	(0.195)	(0.255)	
II	-0.789***	-0.885***	351
	(0.192)	(0.250)	
III	-0.109	-0.491*	351
	(0.194)	(0.297)	
IV	2.514***	3.075***	351
	(0.403)	(0.538)	
Panel B: Public services and accessibility			
Share households with electricity, 2011	7.150***	13.50***	351
	(2.080)	(2.876)	
Share households with tap water, 2011	1.892***	4.092***	351
	(0.576)	(0.800)	
Share households with cars, 2011	10.72***	17.92***	351
	(2.295)	(3.248)	
Urban road length, km, 2019	-0.00136***	0.00100***	351
	(0.000477)	(0.000182)	
District avg. distance to work, km, 2011	0.176*	0.567***	208
	(0.101)	(0.201)	
Panel C: Pre-determined characteristics			
Elevation, 100 m	-0.0200	0.0151	351
	(0.0326)	(0.0569)	
Distance from the coast, km	0.000136	0.000508	351
	(0.000324)	(0.000501)	
Distance from nearest river or lake, km	5.73e-05	0.00649	351
	(0.00510)	(0.00816)	
Distance from nearest mineral deposit, km	-0.000623	-0.00336	351
	(0.00133)	(0.00244)	
Ruggedness, m	0.00143**	0.00207*	351
	(0.000709)	(0.00116)	
Bedrock depth, m	-0.0230	0.00629	351
	(0.0220)	(0.0392)	
Crop suitability, tons per hectare	0.120	-0.406	351
	(0.274)	(0.423)	
Panel D: Non-pre-determined characteristics			
Shape in 1950, km	0.818**	-0.228	351
	(0.324)	(0.808)	
Distance from state headquarters, km	-0.000623	-0.00147*	351
	(0.000562)	(0.000813)	
Distance from district headquarters, km	-0.00988*	-0.00676	351
	(0.00533)	(0.00776)	
Distance from nearest city, km	-0.00342	-0.00513	351
	(0.00210)	(0.00375)	
British direct rule	0.226	0.0105	351
	(0.156)	(0.287)	
State Capital	0.957	4.103***	351
	(1.068)	(1.123)	
Control	Area 2010	Area 1950	

Notes: this table reports pairwise correlations between levels and changes in shape and city attributes. Each row reports a coefficient from an OLS regression of shape in 2010 (col. 1) and the 2010-1950 difference in shape (col. 2) on the attribute indicated in each row, controlling respectively for city area in 2010 (col.1) and in 1950 (col. 2). A description of the variables is provided in Section A.4 in the Appendix. Summary statistics are in Table A3. Robust standard errors in parentheses.*** p<0.01,** p<0.05,* p<0.1.

Table A3: Additional summary statistics

	Obs.	*Mean*	*St. Dev.*	*Min*	*Max*
Elevation, m	351	262.20	230.88	0	1590
Distance from the coast, km	351	400.41	334.06	0.14	1315.88
Distance from nearest river or lake, km	351	15.09	16.47	0.01	102.69
Distance from nearest mineral deposit, km	351	77.21	62.18	0.05	315.18
Ruggedness, m	351	102.39	132.38	0	1000.00
Bedrock depth, m	351	5.74	3.98	1.36	19.77
Crop suitability, tons per hectare	351	1.43	0.33	0.00	1.93
Initial shape, km	351	1.01	0.71	0.35	5.82
British direct rule dummy	351	0.66			
State capital dummy	351	0.05			
Distance from state headquarters, km	351	296.56	190.41	0	998.00
Distance from district headquarters, km	351	15.53	28.25	0	138.00
Distance from nearest city, km	351	40.68	41.75	0	342.00
Share households with electricity, 2011	351	0.94	0.06	0.65	1
Share households with tap water, 2011	351	0.56	0.21	0.06	0.94
Share households with cars, 2011	351	0.08	0.05	0.01	0.29
Urban road length, km, 2019	351	773	2613	3.2	35148
Average distance to work, km, 2011	208	5.73	1.26	2.99	11.34

Notes: this table provides summary statistics for the city-level variables employed in Table A2 and in the robustness checks. These variables are described in Section A.4 in the Appendix.

Table A4: First stage and impact of city shape on population, panel results

	(1)	(2)	(3)	(4)	(5)	(6)
	First Stage		IV	OLS	IV	OLS
Dependent variable:	*Shape, km*	*Log area*	*Log population*			
Potential shape, km	1.397***	0.151***				
	(0.228)	(0.0451)				
Log projected population	-1.188***	0.297**				
	(0.269)	(0.117)				
Shape, km			-0.0975**	0.0250***	-0.107**	0.0247***
			(0.0387)	(0.00788)	(0.0442)	(0.00792)
Log area			0.783***	0.165***	0.827***	0.168***
			(0.182)	(0.0309)	(0.211)	(0.0331)
Observations	6,173	6,173	1,325	1,325	1,135	1,135
AP F stat shape	87.81	87.81	69.36		57.34	
AP F stat area	17.84	17.84	14.12		11.36	
KP F stat	21.13	21.13	16.37		13.29	
City FE	Y	Y	Y	Y	Y	Y
Year FE	Y	Y	Y	Y	Y	Y
Sample	Full	Full	Full	Full	Long diff	Long diff

Notes: this table presents the main results in panel format. Each observation is a city-year. Cols. 1 and 2 are similar to cols. 3 and 4 in Table 2, but employ the full panel of cities. Cols. 3 and 4 present the panel version of cols. 1 and 2 of Table 3, estimated using the full panel of cities. Cols. 5 and 6 repeat the same specifications, but for the panel of 351 cities employed in the main long-differences specification. Cols. 1 and 2 show the first stage, estimated over years 1950 and 1992 through 2010. Cols. 3 through 6 show the IV (odd cols.) and OLS (even cols.) estimates of the impact of city shape (in km) and log city area on log population, using data from Census years 1951, 1991, 2001, and 2011. Angrist-Pischke and Kleibergen-Paap F statistics are reported. All specifications include city and year fixed effects. Standard errors clustered at the city level in parentheses. *** p<0.01,** p<0.05,* p<0.1.

Table A5: Robustness to alternative luminosity thresholds

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	First Stage		IV	OLS	First Stage		IV	OLS
Dependent variable:	*Δ Shape, km*	*Δ Log area*	*Δ Log population*	*Δ Log population*	*Δ Shape, km*	*Δ Log area*	*Δ Log population*	*Δ Log population*
Δ Potential shape, km	1.947***	0.214***			1.806***	0.240***		
	(0.191)	(0.0471)			(0.221)	(0.0523)		
Δ Log projected population	-1.940***	0.118			-1.985***	0.0664		
	(0.469)	(0.123)			(0.455)	(0.144)		
Δ Shape, km			-0.0901**	0.0271***			-0.125**	0.0199**
			(0.0377)	(0.00862)			(0.0543)	(0.00857)
Δ Log area			0.839***	0.197***			0.909***	0.226***
			(0.225)	(0.0349)			(0.255)	(0.0431)
Observations	374	374	374	374	320	320	320	320
AP F stat shape	33.42	33.42	33.42		26.34	26.34	26.34	
AP F stat area	10.83	10.83	10.83		8.42	8.42	8.42	
KP F stat	16.49	16.49	16.49		11.77	11.77	11.77	
Luminosity threshold	30	30	30	30	40	40	40	40
Mean dep var in levels, 2010	5.01	123.00	568,361	568,361	4.40	106.23	708,939	708,939
Mean dep var in levels, 1950	0.98	3.49	98,122	98,122	1.02	3.85	113,205	113,205

Notes: this table presents estimates of the first stage and the impact of shape on population, obtained using different definitions of urban footprints. Cols.1, 2, 5, and 6 report the first stage (analogous to cols. 1 and 2 in Table 2). Cols. 3 and 7 (4 and 8) report the IV (OLS) impact of city shape on population (similar to Table 3). The dependent variables and regressors are all defined as long differences 2010-1950 (2011-1951 for population). The luminosity threshold used to define urban areas is 30 in cols. 1 through 4 and 40 in cols. 5 through 8 (the baseline in the paper is 35). Angrist-Pischke and Kleibergen-Paap F statistics are reported. Robust standard errors in parentheses. *** p<0.01,** p<0.05,* p<0.1.

Table A6: First stage, alternative shape indicators

Shape metric	*A. Remoteness*		*B. Spin*		*C. Range*	
	(1)	(2)	(3)	(4)	(5)	(6)
Dependent variable:	*Δ Shape, km*	*Δ Log area*	*Δ Shape, km²*	*Δ Log area*	*Δ Shape, km*	*Δ Log area*
Δ Potential shape	1.110***	0.178***	1.128***	0.00266	2.698***	0.0998***
	(0.165)	(0.0432)	(0.321)	(0.00189)	(0.368)	(0.0226)
Δ Log projected population	-0.533*	0.232**	-1.032	0.464***	-7.567***	0.0443
	(0.276)	(0.116)	(6.823)	(0.106)	(1.558)	(0.130)
Observations	351	351	351	351	351	351
AP F stat shape	26.74	26.74	27.44	27.44	30.26	30.26
AP F stat area	9.19	9.19	19.57	19.57	10.13	10.13
KP F stat	11.7	11.7	16.62	16.62	13.64	13.64
Avg. shape 1950	0.75		1.06		2.95	
Avg. shape 2010	3.45		25.72		13.57	

Notes: this table presents estimates of the first stage for alternative shape metrics. Odd (even) cols. are analogous to col. 1 (2) in Table 2. The dependent variables and regressors are all defined as long differences 2010-1950. The unit of the shape metrics is km, except for the spin index that is in square km. The shape indexes are discussed in Section A.2 in the Appendix. Remoteness (cols. 1 and 2) is the average distance to the centroid. Spin (cols. 3 and 4) is the average squared length of distances to the centroid. Range (cols. 5 and 6) is the maximum distance between two points on the outline of the city. Angrist-Pischke and Kleibergen-Paap F statistics are reported. Robust standard errors in parentheses. *** p<0.01,** p<0.05,* p<0.1.

Table A7: Impact of city shape on population, robustness to alternative shape indicators

Dependent variable:	*Δ Log population, 2011-1951*					
Shape metric	*A. Remoteness*		*B. Spin*		*C. Range*	
	(1)	(2)	(3)	(4)	(5)	(6)
	IV	OLS	IV	OLS	IV	OLS
Δ Shape	-0.118**	0.0306***	-0.00129	0.000891***	-0.0277**	0.00675***
	(0.0557)	(0.00997)	(0.000810)	(0.000264)	(0.0125)	(0.00220)
Δ Log area	0.817***	0.212***	0.587***	0.243***	0.812***	0.216***
	(0.226)	(0.0339)	(0.128)	(0.0294)	(0.222)	(0.0329)
Observations	351	351	351	351	351	351
AP F stat shape	26.74		27.44		30.26	
AP F stat area	9.19		19.57		10.13	
KP F stat	11.7		16.62		13.64	

Notes: this table presents estimates of the relationship between city shape and population for alternative shape metrics. Odd (even) cols. are analogous to col. 1 (2) in Table 3. The corresponding first stage is reported in Table A6. The regressors are defined as long differences 2010-1950. The unit of the shape metrics is km, except for the spin index that is in square km. Angrist-Pischke and Kleibergen-Paap F statistics are reported. Robust standard errors in parentheses. *** p<0.01,** p<0.05,* p<0.1.

Table A8: Impact of city shape on rents, robustness

Dependent variable:	*Δ Log rent 2008-2006, excluding bottom 25%*					
	(1) IV	(2) OLS	(3) IV	(4) OLS	(5) IV	(6) OLS
	All districts		*Only districts with one city*		*Only top city per district*	
Δ Shape, km	-0.663	0.00421	-0.532	-0.00729	-0.769	0.0188
	(0.557)	(0.0487)	(0.333)	(0.0692)	(0.704)	(0.0493)
Δ Log area	-2.535	-0.0125	-1.354	-0.103	-2.138	-0.0574
	(2.300)	(0.0927)	(1.131)	(0.112)	(2.282)	(0.100)
Observations	262	262	134	134	215	215
AP F stat shape	9.60		14.77		5.11	
AP F stat area	3.00		6.12		2.80	
KP F stat	1.67		2.93		1.20	

Notes: this table is analogous to Table 5, but the district averages of rents exclude the bottom 25% of the rents distribution. Robust standard errors in parentheses. *** p<0.01,** p<0.05,* p<0.1.

Table A9: Impact of city shape on rents, IHDS data

Dependent variable:	*Δ Log rent 2010-2005*		*Δ Log rent residual 2010-2005*	
	(1) IV	(2) OLS	(3) IV	(4) OLS
Δ Shape, km	-0.0624	0.0121	-0.0382	0.0250
	(0.0761)	(0.0164)	(0.0789)	(0.0190)
Δ Log area	-0.292	0.203*	-0.279	0.193+
	(0.736)	(0.116)	(0.766)	(0.128)
Observations	111	111	111	111
AP F stat shape	6.85		6.85	
AP F stat area	4.11		4.11	
KP F stat	2.25		2.25	
Source	IHDS	IHDS	IHDS	IHDS

Notes: this table is analogous to Table 5, cols.1 and 2, but uses rents from a different source. Each observation is a district. In cols. 1 and 2 the dependent variable is the 2010-2005 long difference in log monthly total rents, averaged at the district level, from the IHDS dataset. In cols. 3 and 4 the dependent variable is the long difference in the log average rent residual, from a hedonic regression of rent on housing characteristics discussed in Section A in the Appendix. Robust standard errors in parentheses. *** p<0.01,** p<0.05,* p<0.1.

Table A10: Pairwise correlations between instruments and city characteristics

Dependent variable:	(1) *Δ Potential shape, km, 2010-1950*	(2) *Δ Log projected population, 2010-1950*
Elevation, 100 m	-0.0245	0.0127
	(0.0396)	(0.0121)
Distance from the coast, km	-0.000303	-0.000242***
	(0.000262)	(9.16e-05)
Distance from nearest river or lake, km	-0.00307	0.00340
	(0.00473)	(0.00212)
Distance from nearest mineral deposit, km	-0.00301*	-0.000135
	(0.00169)	(0.000675)
Ruggedness, m	0.000699	0.000368*
	(0.000507)	(0.000221)
Bedrock depth, m	0.00455	-0.0113*
	(0.0232)	(0.00633)
Crop suitability	0.299	0.247*
	(0.306)	(0.135)
Observations	351	351

Notes: this table reports estimates of the relationship between the instruments and time-invariant city characteristics. Each row reports a coefficient from an OLS regression of the 2010-1950 long differences in potential shape (col. 1) and log projected population (col. 2) on the controls indicated in each row. The controls are described in Section A.4 in the Appendix. Robust standard errors in parentheses.*** p<0.01,** p<0.05,* p<0.1.

Table A11: IV impact of city shape on population, robustness to sample cuts

	Dependent variable: Δ Log population, 2011-1951						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Δ Shape, km	-0.104**	-0.0999**	-0.155*	-0.175**	-0.0857**	-0.0932**	-0.0854**
	(0.0475)	(0.0486)	(0.0800)	(0.0748)	(0.0412)	(0.0422)	(0.0408)
Δ Log area	0.856***	0.863***	1.103***	0.961***	0.804***	0.818***	0.812***
	(0.246)	(0.265)	(0.415)	(0.321)	(0.234)	(0.220)	(0.233)
Observations	337	334	242	204	318	316	316
AP F stat shape	29.02	24.16	8.91	9.22	27.33	28.51	27.27
AP F stat area	8.72	7.76	4.13	4.03	9.14	10.13	8.62
KP F stat	12.36	10.92	6.97	9.96	12.18	14.09	12.10
Excluded cities	Mountainous	Coastal	River/lake	Mineral	Top 90% ruggedness	Top 90% bedrock depth	Top 90% crop suitability

Notes: this table reports the same IV specification as in Table 3, col.1, for various sample cuts discussed in Section 7. A description of the controls is provided in Section A.4 in the Appendix. Robust standard errors in parentheses.*** p<0.01,** p<0.05,* p<0.1.

Table A12: Robustness to confounding trends, non-predetermined characteristics

Characteristic:	*A. Initial shape*			*B. British direct rule*			*C. State capital*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	First Stage		IV	First Stage		IV	First Stage		IV
Dependent variable:	*Δ Shape, km*	*Δ Log area*	*Δ Log population*	*Δ Shape, km*	*Δ Log area*	*Δ Log population*	*Δ Shape, km*	*Δ Log area*	*Δ Log population*
Δ Potential shape, km	1.566***	0.319***		1.943***	0.239***		1.700***	0.222***	
	(0.203)	(0.0480)		(0.250)	(0.0490)		(0.186)	(0.0489)	
Δ Log projected population	-1.542***	-0.112		-2.228***	0.0389		-2.025***	0.0555	
	(0.439)	(0.127)		(0.486)	(0.132)		(0.396)	(0.132)	
Δ Shape, km			-0.271**			-0.0963**			-0.111***
			(0.116)			(0.0446)			(0.0418)
Δ Log area			1.337***			0.851***			0.814***
			(0.456)			(0.238)			(0.224)
Control	1.764***	-0.409***	1.060**	-0.0654	-0.221**	-0.0109	6.228***	0.271	0.940***
	(0.435)	(0.111)	(0.468)	(0.259)	(0.0904)	(0.0915)	(1.173)	(0.209)	(0.285)
Observations	351	351	351	351	351	351	351	351	351
AP F stat shape	6.09	6.09	6.09	26.33	26.33	26.33	33.92	33.92	33.92
AP F stat area	3.96	3.96	3.96	8.85	8.85	8.85	9.55	9.55	9.55
KP F stat	10.02	10.02	10.02	12.78	12.78	12.78	14.38	14.38	14.38

Notes: this table extends the robustness checks of Table 6, showing estimates of the first stage and of the impact of shape on population, controlling for initial shape (cols. 1 through 3), a British direct rule dummy (cols. 4 through 6), and a state capital dummy (cols. 7 through 9). Cols.1, 2, 4, 5, 7, and 8 report the first stage (analogous to cols. 1 and 2 in Table 2). Cols. 3, 6, and 9 report the IV impact of city shape on population (similar to Table 3, col.1). The dependent variables and regressors are all defined as long differences 2010-1950 (2011-1951 for population). Angrist-Pischke and Kleibergen-Paap F statistics are reported. Robust standard errors in parentheses. *** p<0.01,** p<0.05,* p<0.1.

Table A13: Robustness to sample cuts, non-predetermined characteristics

Excluded cities:	***A. Shrinking***			***B. Fast growing***			***C. Slow growing***			***D. Constrained***		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	First Stage		IV	First Stage		IV	First Stage		IV	First Stage		IV
Dependent variable:	*Δ Shape*	*Δ Log area*	*Δ Log population*	*Δ Shape*	*Δ Log area*	*Δ Log population*	*Δ Shape*	*Δ Log area*	*Δ Log population*	*Δ Shape*	*Δ Log area*	*Δ Log population*
Δ Potential shape, km	1.945***	0.238***		1.669***	0.232***		1.931***	0.225***		1.861***	0.239***	
	(0.250)	(0.0495)		(0.171)	(0.0542)		(0.250)	(0.0484)		(0.257)	(0.0553)	
Δ Log projected population	-2.255***	0.00689		-2.214***	-0.0382		-2.242***	0.0118		-2.214***	0.00894	
	(0.507)	(0.135)		(0.412)	(0.147)		(0.511)	(0.134)		(0.492)	(0.149)	
Δ Shape, km			-0.126**			-0.0728*			-0.124**			-0.125**
			(0.0510)			(0.0426)			(0.0503)			(0.0567)
Δ Log area			1.034***			0.578**			1.028***			0.969***
			(0.265)			(0.227)			(0.275)			(0.268)
Observations	336	336	336	316	316	316	316	316	316	316	316	316
AP F stat shape	21.25	21.25	21.25	25.12	25.12	25.12	21.12	21.12	21.12	22.52	22.52	22.52
AP F stat area	7.17	7.17	7.17	6.13	6.13	6.13	6.64	6.64	6.64	6.83	6.83	6.83
KP F stat	10.49	10.49	10.49	9.38	9.38	9.38	9.74	9.74	9.74	9.84	9.84	9.84

Notes: this table extends the robustness checks of Table A11, showing estimates of the first stage and of the impact of shape on population, excluding particular sets of cities, discussed in Section 7. Cols.1, 2, 4, 5, 7, 8, 10, and 11 report the first stage (analogous to cols. 1 and 2 in Table 2). Cols. 3, 6, 9, and 12 report the IV impact of city shape on population (similar to col. 1 in Table 3). The dependent variables and regressors are all defined as long differences 2010-1950 (2011-1951 for population). Angrist-Pischke and Kleibergen-Paap F statistics are reported. Robust standard errors in parentheses. *** p<0.01,** p<0.05,* p<0.1.

Table A14: Falsification test with lagged outcomes, wages and rents

Dependent variable	*Δ Log rents, 2008-2006*	*Δ Log rents, 2008-2006*	*Δ Log wages, 1995-1992*	*Δ Log wages, 1998-1994*	*Δ Log wages, 1995-1992*	*Δ Log wages, 1998-1994*
	(1)	(2)	(3)	(4)	(5)	(6)
Potential shape, km, 2000					-0.000151	
					(0.210)	
Potential shape, km, 2005					-0.0549	
					(0.217)	
Potential shape, km, 2007						-0.221*
						(0.125)
Potential shape, km, 2008	-0.0117					
	(0.166)					
Potential shape, km, 2010	-0.00202					0.120
	(0.132)					(0.0988)
Δ Potential shape, km, 2005-2000			0.00863			
			(0.196)			
Δ Potential shape, km, 2010-2005				-0.0465		
				(0.0339)		
Δ Potential shape, km, 2010-2009		0.0139				
		(0.0697)				
Projected population, 2000					-0.304	
					(0.691)	
Projected population, 2005					0.379	
					(0.669)	
Projected population, 2007						0.0337
						(0.910)
Projected population, 2008	3.771**					
	(1.709)					
Projected population, 2010	-3.731**					0.0464
	(1.701)					(0.914)
Δ Projected population, 2005-2000			0.290			
			(0.490)			
Δ Projected population, 2010-2005				0.500		
				(0.553)		
Δ Projected population, 2010-2009		-6.622*				
		(3.409)				
Observations	303	303	168	191	168	191

Notes: this table presents a falsification test similar to that of Table 7, to show that the instrument is not correlated with past changes in rents and wages. The dependent variables are defined including all districts. Robust standard errors in parentheses. *** p<0.01,** p<0.05,* p<0.1.

Table A15: Heterogeneous effects of transit, robustness

	Dependent variable: Δ Log population, 2011-1951							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Δ Shape, km	-0.248**	-0.432**	-0.193**	-0.224**	-0.195**	-0.315**	-0.296**	-0.252**
	(0.115)	(0.219)	(0.0925)	(0.0886)	(0.0822)	(0.125)	(0.117)	(0.111)
Δ Shape · Transit	7.48e-06**	0.000170**	0.0108**	0.321***	0.407*	0.000356**	0.000775***	0.000190*
	(3.44e-06)	(8.07e-05)	(0.00474)	(0.121)	(0.227)	(0.000141)	(0.000296)	(0.000103)
Δ Log area	1.195***	1.517**	1.044***	1.050***	0.973***	1.480***	1.437***	1.338***
	(0.459)	(0.699)	(0.369)	(0.329)	(0.327)	(0.521)	(0.493)	(0.445)
Observations	336	336	336	123	123	246	246	246
AP F stat interaction	1079.77	35.7	358.48	37.59	31.18	644.28	693.9	328.63
AP F stat shape	11.04	5.04	11.6	5.61	6.60	6.98	7.62	7.38
AP F stat area	5.13	3.18	6.93	9.55	12.39	3.63	3.84	4.35
KP F stat	8.19	5.57	9.74	8.66	8.06	6.36	6.61	6.95
Interaction variable	Roads 2019	Roads 1981	State roads 1981	Proximity	Grid roads	Cars 2011	Cars 2001	State cars 1984

Notes: this table reports the same specifications of Table 9, but additionally controls for the number of banks in 1981. Robust standard errors in parentheses.*** p<0.01,** p<0.05,* p<0.1.

Table A16: Impact of city shape on infrastructure and transit

Panel A: Roads

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	IV	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS	IV	OLS
Dependent variable:	*Log roads*		*Log motorways*		*Log roads per capita*		*Log motorways per capita*		*Δ Log roads*		*Δ Log roads per capita*	
Shape, km	-0.161**	-0.0138	-0.707*	-0.490***	-0.0683	-0.0923***	-0.615*	-0.568***				
	(0.0728)	(0.0133)	(0.364)	(0.171)	(0.0507)	(0.0202)	(0.343)	(0.178)				
Log area	1.823***	1.215***	5.047***	4.416***	0.419**	0.639***	3.642**	3.840***				
	(0.237)	(0.0497)	(1.553)	(0.838)	(0.188)	(0.0745)	(1.503)	(0.856)				
Δ Shape, km									-0.0236	0.0815***	-0.0180	0.0659**
									(0.0529)	(0.0296)	(0.0536)	(0.0298)
Δ Log area									0.626	0.397***	0.377	0.277***
									(0.398)	(0.0917)	(0.456)	(0.102)
Observations	351	351	351	351	351	351	351	351	335	335	335	335
AP F stat shape	8.44		8.44		8.44		8.44		32.68		32.68	
AP F stat area	25.08		25.08		25.08		25.08		11.32		11.32	
KP F stat	6.73		6.73		6.73		6.73		15.69		15.69	

Panel B: Akbar et al. (2019)

	(1)	(2)	(3)	(4)	(5)	(6)
	IV	OLS	IV	OLS	IV	OLS
Dependent variable:	*Grid conformity*		*Mobility*		*Proximity*	
Shape, km	0.0149**	0.00519**	-0.0352*	-0.0165***	-0.0172	0.0158***
	(0.00639)	(0.00251)	(0.0202)	(0.00461)	(0.0210)	(0.00468)
Log area	-0.0940**	-0.0325**	0.130	0.0529**	0.174	-0.0310
	(0.0369)	(0.0131)	(0.0969)	(0.0238)	(0.117)	(0.0296)
Observations	128	128	128	128	128	128
AP F stat shape	1.78		1.78		1.78	
AP F stat area	4.38		4.38		4.38	
KP F stat	2.76		2.76		2.76	

Notes: this table reports estimates of the impact of city shape on infrastructure-related variables described in Section 9 and Section A.5 in the Appendix. In all cols. other than cols. 9 through 12, the regressors are city shape, in km, and log city area, measured in 2010. Panel A, cols. 1 through 8 considers the length of roads (motorways) in a city's 2010 lit-up shape, as reported in 2019 in Openstreetmap. In cols. 5 through 8 the dependent variable is the log of road length normalized by city population ins 2011. In cols. 9 through 12 the dependent variable is the log difference of 2019 Openstreetmap roads and 1981 city roads (from the Census) and the regressor s are 2010-1950 changes in city shape and log city area. Panel B considers indexes from Akbar et al. (2019), measured in 2016. Grid conformity (cols. 1 and 2) is a measure the regularity of a city's primary road grid. Mobility (col. 3 and 4) is a speed-based index of vehicular mobility. Proximity (cols. 5 and 6) is an index of distance accessibility from Akbar et al (2018). Means of the dependent variables are reported in Table 9. Estimation is by IV in odd columns, and OLS in even columns. Angrist-Pischke and Kleibergen-Paap F statistics are reported. Robust standard errors in parentheses.*** p<0.01,** p<0.05,* p<0.1.