WPS8678

Policy Research Working Paper 8678

Measuring Urban Economic Density

J. Vernon Henderson
Dzhamilya Nigmatulina
Sebastian Kriticos

WORLD BANK GROUP
Development Economics
Development Research Group
December 2018

Policy Research Working Paper 8678

Abstract

Agglomeration economies are at the heart of urban economics, driving the existence and extent of cities and are central to structural transformation and the urbanization process. This paper evaluates the use of different measures of economic density in assessing urban agglomeration effects, by examining how well they explain household income differences across cities and neighborhoods in six African countries. The paper examines simple scale and density measures and more nuanced ones that capture the extent of clustering within cities. The evidence suggests that more nuanced measures attempting to capture within-city differences in the extent of clustering do no better than a simple density measure in explaining income differences across cities, at least for the current degree of accuracy in measuring clustering. However, simple city scale measures, such as total population, are inferior to density measures and to some degree misleading. The analysis finds large household income premiums from being in bigger and particularly denser cities over rural areas in Africa, indicating that migration pull forces remain very strong in the structural transformation process. Moreover, the marginal effects of increases in urban density on household income are very large, with density elasticities of 0.6. In addition to strong city-level density effects, the analysis finds strong neighborhood effects. For household incomes, overall city density and density of the neighborhood matter.

This paper is a product of the Development Research Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/research. The authors may be contacted at J.V.Henderson@lse.ac.uk.

Measuring Urban Economic Density*

J. Vernon Henderson† Dzhamilya Nigmatulina Sebastian Kriticos

Keywords: Economic density, agglomeration, scale economies, city definitions, Africa, economic development, Landscan, land use.

JEL: J24, J31, O15, O55, R1, R3

*This work was supported by the World Banks Strategic Research Program funded by the UK Department for International Development and by an Africa Research Program on Spatial Development of Cities at LSE and Oxford which is funded by the Multi Donor Trust Fund on Sustainable Urbanization of the World Bank, and by the UK Department for International Development. We thank Patricia Jones for sharing value added, location, employment and industrial code data for Kampala firms from the Uganda Business Inquiry survey of 2002.

†**Corresponding author:** *J.V.Henderson@lse.ac.uk*. All authors are at LSE.

Introduction

At the heart of urban economics are the sources and nature of agglomeration economies, which drive the existence and extent of cities. Agglomeration economies are also central to structural transformation in developing countries: why people urbanize. The literature is replete with studies attempting to estimate the productivity or wage premium from being in cities compared to rural areas or from being in bigger versus smaller cities (e.g. Ciccone and Hall (1996) and Glaeser and Mare (2001), with reviews of the literature in Combes and Gobillon (2015) and Rosenthal and Strange (2004)). These studies adopt specific, simple measures of agglomeration such as indicator variables for settlement type (e.g. urban versus rural), a continuous total population measure, or at best, a basic population density measure, where the denominator (area) is often inconsistently defined across cities. They also adopt a specific definition of cities and urbanized areas, often chosen by national statistics bureaus based on qualitative aspects of land use and the built environment, degree of centrality of activity, and other often less than precise measures. In some countries, population density plays a role in the definition but usually not a central one.

In general, these studies do not provide any statistical rationale for their choices. For example, city definitions are driven by the convenience of adopting some official definition. Measures of agglomeration are typically chosen with little analysis as to why, with no quantitative evaluation of what measure(s) would be most relevant. This paper will focus on an evaluation of different agglomeration measures which we characterize as economic density measures. We first consider traditional options: urban versus rural, a continuous total population measure, and a continuous measure of overall population density. However, these measures do not capture, for a given population and overall density, the degree to which economic activity is clustered within a city. Two cities with the same density and population may have very different levels of clustering of economic activity within the city, which can be captured by other measures reflecting variation in population densities within a city. We derive and utilize several measures. This issue is important given the increasing use of the De La Roca and Puga (2017) measure (for example, Collier et al. (2018)), which reflects to some extent the degree of clustering. Does simple population density suffice or do we learn more by using more nuanced measures capturing aspects of clustering?

To evaluate the efficacy of different measures, consistent with the literature, the paper will examine how well different measures explain income differentials across space. We do this for a set of six African countries in a sample that covers the rural sector, 193 low-density urban settlements with over 5,000 people, and 115 high-density cities with over 50,000 people. We will adopt consistent definitions of urbanized areas. These consistent definitions will be based on population densities for each 1 km grid square across space. We aggregate contiguous squares of high density to create cities – which are the consolidation of an urban core and a surrounding lower density fringe – and aggregate contiguous squares of lower density to create stand alone, low density [LD] settlements. We do not know the "best" definitions but we will use ones which are consistent across countries and accord with other population density based initiatives (for example, OECD (2012)). While admittedly the density and population thresholds we choose are to some degree arbitrary, they are based on types of thresholds some countries and researchers cite or use – other papers tackle that problem more directly.

For the measurement of economic density, there is the issue of determining relevant spatial scales. While most papers explore agglomeration benefits at the level of a city or county, a few papers look within cities at the extent of spatial decay, finding a very rapid spatial decay (Arzaghi and Henderson (2008), Rosenthal and Strange (2008)). In Arzaghi and Henderson (2008) spatial decay of estimated effects within New York City is so quick as to beg the question of why such a huge city even exists. Clearly, there must be some other overall benefits, such as labor market externalities or greater input varieties from being in New York beyond density in the own neighborhood. Here we explore both effects together: those from overall city economic density and from own local neighborhood economic density. We will ask also if people living in the urban core versus fringe of the city benefit differentially from overall density characteristics of the city.

We also explore some economic issues. How do marginal scale effects, or elasticities, based on correlations with income measures vary across the spatial hierarchy: rural areas, LD settlements, and cities? Do scale effects vary by the way income is measured: personal income from all sources or wage income, as compared to household income? What are the issues in looking at each income measure and what household or personal characteristics

should be controlled for?

Our work is based on a set of developing countries in Africa where there is a literature focused on push-pull issues in the urbanization process. These are reviewed in Henderson and Kriticos (2018) or Gollin et al. (2016), with analysis of classical push arguments in Schultz et al. (1968), Matsuyama (1992), Gollin et al. (2007), and Bustos et al. (2016) and an analysis of pull arguments in Lewis (1954), Hansen and Prescott (2002), and Galor and Mountford (2008). While we know climate deterioration or conflict may also push people into cities (Henderson et al. (2017), Fay and Opal (1999), Barrios et al. (2006), Brückner (2012)), we want to see if in Africa there is the appearance of strong income pull factors. That is motivated by a literature which argues that, while African cities have high population densities, they remain unproductive for the development of traded goods such as manufactured goods because they have low economic density (Collier and Jones (2016), Venables (2018), and Lall et al. (2017)). Low economic density arises if firms and economic activity more generally are not clustered but spread throughout the city, potentially because of the high costs of commuting within cities inducing firms to locate nearer to residents (Fujita and Ogawa, 1982). We will explore whether African cities have a lower degree of clustering of economic activity relative to the rest of the world and we will take a detailed look at patterns of clustering within Nairobi and Kampala.

In the paper, Section 1 starts with how we define cities and what are the advantages and disadvantages of the LandScan (2012) dataset on economic activity that we use. We then try to ground-truth LandScan (2012) data using Nairobi and Kampala as test cases. Section 2 defines various measures of economic density for urbanized and rural areas, decomposing them into first and second moment components. Section 3 compares the extent of economic density and its attributes for African cities versus other cities worldwide. Section 4 looks at the relationship between different measures of economic density and income differentials across the whole spatial hierarchy. Section 5 looks specifically at cities and examines issues such as the optimal rate of spatial discount for De La Roca and Puga (2017) measures of neighbor effects; the role of the second moment aspects of economic density measures; and how important local density measures are within cities, as well as location within the city. Section 6 looks in detail at how the productivity of firms within Kampala is related to the

characteristics of the neighborhood in which they locate.

1 Using Landscan Data and Defining Urbanized Areas

1.1 Landscan data

To analyze measures of economic density, we need fine spatial resolution data to calculate neighborhood effects and variation in clustering within a city. And we need the most accurate data available. For countries like the USA, both finely gridded population and employment data are available from censuses. However, in most developing countries that is not the case. Population data are only available at a coarse scale such as regional or local government political unit, and economic censuses in Sub-Saharan Africa are generally non-existent. Even when they do exist (e.g. Uganda), they tend not to be publicly available.

Our primary data source is LandScan (2012) from Oak Ridge National Laboratory in the USA, which is now being used in some research (e.g. Desmet et al. (2018)). Oak Ridge takes population data from censuses and other sources worldwide on as fine a spatial scale for each country as they can obtain. They then create a measure of an ambient population for each 1km grid square on the planet, in a process we describe below. The ambient population is meant to represent where people are on average over the 24 hour day. To assess the ambient population, they appear to use nocturnal and diurnal population estimates for at least some areas of the globe, although these are not publicly available. Later in a ground-truthing exercise for Nairobi and Kampala, we will demonstrate our own interpretation of how nocturnal and diurnal populations might be estimated and combined.

For Landscan, as for WorldPop[1], the Global Human Settlements Layer[2], and similar data sets, a key element in this process involves taking population numbers at some upper level of spatial scale and allocating people to fine grid squares based on where they are likely to live and possibly work. The typical standard in such work has been to allocate people on the basis of the relative extent of ground cover in a grid square from Landsat satellite imagery, or its enhanced versions.

[1] http://www.worldpop.org.uk/data
[2] http://ghsl.jrc.ec.europa.eu/datasets.php

Landscan has two key advantages and two key disadvantages over other data sets. First, Oak Ridge National Lab is more explicit in the fact that they are trying to estimate the ambient population with potentially nocturnal and diurnal populations; while in other algorithms this is implicit through the general smearing of the population into built cover, without workplace or residence distinction. The second advantage of Landscan is that Oak Ridge National Lab has access to information which would improve precision over the use of Landsat to just assign people to built cover. Oak Ridge has access to very high-resolution satellite data (under 10-40cm) which *potentially* allows them to distinguish building types based on what building shapes are likely to house employment versus residents (versus commercial activities like shopping), as well as *potentially* to distinguish roads for commuting and even infer building heights with digital elevation modeling [DEM]. A key disadvantage in using Landscan is the complete lack of specificity and transparency as to what Oak Ridge researchers actually do; hence, our use of the word "potentially" in describing what they might do. The second disadvantage, which they acknowledge, is that Landscan data for different time periods are not comparable over time, presumably both because of differential availability of high-resolution data over time and increasingly sophisticated extraction of information from later satellite images.

Hopefully in the future, proposed data sets such as the High Resolution Human Settlement Layer (HRSL)[3] or Modelling and Forecasting African Urban Population Patterns (MAUPP)[4] – which also use very high spatial resolution data – will be able to cover a wider set of time periods and countries consistently with a more explicit methodology. Then one will be able to compare them with Landscan and do more comprehensive ground-truthing exercises. Second while ambient population may be a relevant measure to use in characterizing economic density, one might prefer to know about clustering and density of employment. If the assignment of people to workplace buildings gets very sophisticated, we would be able to explore the use of employment density measures, as well as ambient population ones.

[3]https://www.ciesin.columbia.edu/data/hrsl
[4]http://spell.ulb.be/project/maupp

1.2 Ground-truthing Landscan

Here we attempt to ground-truth Landscan measures at the grid square level for Kampala and Nairobi, two cities where we have fine spatial resolution data on population and employment which is unavailable to Oak Ridge National Lab. We further attempt to replicate Landscan's ambient population measure by grid square using an assignment algorithm on our own data. We will conclude that Landscan measures do well and seem superior to other commonly used measures which smear population into grid squares on the basis of built cover on the ground.

In the upper panels of Figures 1 and 2, we show the population and employment distribution for Kampala and Nairobi in 1 km grid squares. The population data for Nairobi are at the level of 2,213 enumeration units for 2009 contained in the 2015 built area of Nairobi defined in Henderson et al. (2018). For Kampala in 2002, population is at the level of 174 parishes within the administrative unit of Greater Kampala. We assign population levels from these survey units to the 1km grid square level by applying a weighted sum to the survey area numbers, where the weights reflect the share of land mass from each survey area(s) that falls within a 1km grid square. To make Kampala 2002 population comparable to 2011 employment numbers, we blow up the population in each grid square by an overall population growth rate of 3% per annum from 2002 to 2011. For employment, we use the economic census, which covers private and public employment for Kampala for 2011, and provides exact location points of firms across the city. One issue is that total employment in the census is far below known estimates; hence, given the age distribution in Kampala and labor force participation of urban Uganda, we have multiplied each grid squares employment by 2.761 to make up for the employment deficit.[5] The implicit assumptions in allocating growth and under-counting of employment to grid squares are obvious.

For Nairobi in 2009, we can quite accurately infer population of the grid square, based on fine-scale EA data. However, we do not know total employment, nor its distribution. We infer total employment based on Nairobi's 2009 population, and labor force participation and age distribution numbers for urban Kenya. Since there is no economic census for

[5]The World Bank estimates that labor force participation of people aged 15 or more in Uganda is 0.71. There are 1,704,604 people of age 15+ in Kampala from the 2011 census. Thus, approximately 1,210,267 people should work out of the total city population of 2,957,505. The economic census only captures 438,374 of these.

Nairobi, we obtain the distribution of employment using data from Henderson et al. (2018), where for each grid square we know the footprint and height of every building in Nairobi in 2015 – from aerial photo and Lidar data – and can calculate building volume. We match these buildings with land use maps, before taking total employment of the city and smearing it into grid squares according to each grid squares share in total volume of non-residential buildings in Nairobi. The alternative would be to smear it into commercial and industrial buildings, ignoring public buildings. Crucially, unlike Landscan we do not need to base smearing on inferences from satellite images of what uses buildings have; instead, we know the use and the volume fairly accurately.

For each city we create a measure of the ambient population according to the following equation:

$$Replication_i = \left(\frac{10}{24}\right) Emp_i + \left(\frac{14}{24}\right) Pop_i + \left(\frac{10}{24}\right)(1 - LFP_c) Pop_i \tag{1}$$

We base our replication of the ambient population just on places of work and residence, where we assume for 14 hours of a day (nocturnal) all people are in their grid square of residence to sleep, eat, and recreate. For 10 hours a day, we add in the employment in the grid square, allowing people time to work, hangout, and finish commuting. We then add in the non-working population of the grid square assuming they remain in that square kilometer and subtract out the resident workers (since we have already counted employment). If everyone works in their grid, then we just have total grid square population; but, for downtown grid squares where few people live, most have replication numbers from employment. We make no allowance for the time people are on roads or shopping outside the grid square. We have no information on which to base such inferences, especially in a context where so many people commute by walking.

In Figure 1 for Kampala, we show 4 items. As noted above, in the upper left panel of each figure is the population distribution over space and on the right upper panel is the employment distribution. In the bottom panel, on the left, we have Landscan numbers; and on the right, we have our replication numbers. For Kampala, we see the overall monotonicity of the city. Although it is hard to see, the very low bar population grid squares near the center are to some degree filled in by where employment spikes. The bottom

right panel shows our smearing to get the ambient population. The Landscan figure has an obvious degree of smoothing, with reduced peak heights and assignment of lots of people into low-density grid squares. Of course, it could be that Landscan is allocating people during commuting times to roads and to shopping areas, and that is the reason for the smoothing. Overall it seems Landscan may do a reasonable job: the simple correlation coefficients of Landscan numbers with population, employment and our replication numbers are respectively 0.55, 0.60, and 0.60.

For Nairobi in Figure 2, we note our employment patterns lack the sharp peaks of Kampala, in part because we smear employment into non-residential buildings, including public buildings. If we just smeared into commercial and industrial buildings we would get sharper peaks near the center, but that does not mean it is a better choice. The Landscan figure again exhibits a degree of smoothing, missing the sharper peaks we see in our replication, as well as missing high-density slum areas to the south-west of the city center. However, Landscan does seem to do a better job of capturing low-density grid squares near the city center in Nairobi than it does for Kampala. For Nairobi, the simple correlation coefficients of Landscan numbers with population, employment and our replication numbers are higher at respectively 0.65, 0.56 and 0.69.

For Nairobi, we also know exactly the footprint (or ground cover) of all buildings. We can calculate what would happen if we just smeared total urban area population by share of each grid square in total urban area built cover, hence replicating what Landsat-based smearing exercises seem to be trying to do. The bottom panel of the figure shows the result. Inferred density is basically flat throughout the city. As in Henderson et al. (2018), the figure implies built cover per grid square is basically flat throughout the city (while building volume and height decline sharply with distance from the center). This clearly shows that smearing population into built cover to calculate within-city density variation would be more problematic than using Landscan data.

1.3 Defining Urbanized Areas

As noted earlier, the problem with typical definitions of urbanized areas – from the United Nations or economic censuses of different countries for instance – is that they employ

country-specific city and settlement definitions, which means there is no consistency across countries. Second, although occasionally definitions are somewhat density based, most defintions are based on qualitative and subjective criteria including governance elements. Many countries define cities based on status in the political-spatial hierarchy, local political boundaries defined historically, or through an application or evaluation process to redefine rural areas as cities, which tends to under-represent newer fast-growing agglomerations due to delays both in application and evaluation as well as the granting of status.

We employ a consistent density based definition across our African countries, using LandScan (2012) population per grid square. We set population per grid square, or density thresholds to define cities and settlements. To do so, we apply a smoothing algorithm so that each own grid square is assigned the average density of neighbors within 7km. Smoothing is essential to avoid large doughnut holes in cities, due to terrain factors, airfields, parks, big open public spaces and the like. We define a core city as a set of contiguous grid squares all of which have a density greater than or equal to 1,500 per sq. km. and the population of these contiguous squares must sum to 50,000 or more. The area included in these contiguous squares over 1,500 per sq. km. defines the area and population of what we call the city core. We then add in a fringe to each city core, which includes all contiguous grid squares with population density over 500 per sq. km. The core combined with a fringe is called a city.

For smaller urbanized places that are stand-alone, we require a collection of contiguous grid squares all with (a smoothed) population density over 500 per sq. km., which collectively sum to 5,000 or more. We call these low density [LD] settlements. Full details of these urban definitions are given in Appendix B.

The process and impact of threshold decisions are illustrated in Figure 3 for Nairobi. Core city areas are in dark blue, and overall cities are also outlined in dark blue. There are two cities in the figure, Machakos to the bottom right and Nairobi. Nairobi consists of the main core and three small core areas, essentially satellite towns now falling under the umbrella of Nairobi. The fringe of Nairobi consists of pink and light blue areas, within the dark blue outline. Our choice of 500 per sq. km. is based on the idea that a lower threshold such as 300 per sq. km. (yellow areas) is too loose and extends too far into more rural

and low-density settlement areas much further north of Nairobi. And it would place the center of Nairobi well outside its true central core. A higher cutoff of, say, 750 people per sq. km. (light blue areas) may be too stringent and exclude satellite cities around Nairobi that are very likely to be within the commuting zone. Obviously, other arguments about drawing boundaries can be made. In the figure we also outline in green the independent LD settlements. Some are very spatially distinct, but some follow ribbons (roads) to the north outside Nairobi, where rural areas are interspersed with urbanized settlements. In the figure everything in yellow or the Google Earth background is rural.

2 Defining Economic Density

Figure 4 illustrates issues about urban density definitions.[6] All hypothetical cities in Figure 4 have the same total population (180) in thousands and average density (5). City 1 has no clustering. Cities 2 and 3 have the same degree of within grid square clustering, with half the grid squares with no population and half with 10 people per grid. The 10 means greater within grid square possibilities for intersecting with others (the pairwise possibilities for meetings for example, ((n-1)!). However city 2 allows for more possibilities for interactions with neighbors. Ignoring the boundaries in city 3, on average a grid squares has 40 queen neighbors, while in city 2 a grid square has 80 queen neighbors.

We now turn to two measures which reflect these differences, personal population density and De La Roca and Puga (2017) density [RPA]. For a given city area, personal population density is a weighted, rather than simple, sum of own cell population densities. So in Figure 4, that gives a value of 5 for city 1 and 10 for cities 2 and 3. The RPA measure further makes a distinction between cities 2 and 3. It does a sum of grid square measures, where each grid square measure is a distance discounted sum of your own and neighbors density out to a given radius. Each grid square measure is weighted by its population share in the city. This measure will give a higher value for city 2 than 3.

The basics of what we present is not our invention. Modi (2004) proposed the idea and term personal population density. Small and Cohen (2004) calculate, on a coarser scale,

[6] Figure 4 and the definition and decomposition of personal population density are borrowed from on-going work by Henderson, Storeygard and Weil. This is gratefully acknowledged.

a spatial Gini as a measure of within-gridcell variation in activity. De La Roca and Puga (2017) calculate the RPA measure we use, based on the city 2 idea that neighbors matter. What we add, based on on-going work by Vernon Henderson, Adam Storeygard and David Weil, is a decomposition for personal population density, which as far as we know is new; and, in this paper, we do a similar one for the RPA measure. Also De La Roca and Puga (2017) do not apply a distance discount factor. Below, we experiment empirically to try to find the discount rate that optimizes the added explanatory power of the economic density measures.

For personal population density [PPD] the measure for city j with N_j cells is:

$$PPD_j = \sum_i^{N_j} P_{ij} \frac{P_{ij}}{P_j} = PD_j \left(1 + \frac{Var(P_{ij})}{PD_j^2}\right) = PD_j(1 + CV(P_{ij})^2) \tag{2}$$

where CV: coefficient of variation; N_j: number of grid sqs.; and $PD_j = \frac{\sum_i^{N_j} P_{ij}}{N_j}$.

PPD can be decomposed into overall population density [PD], a typical scale measure, and one plus the coefficient of variation. The latter captures the degree of variation relative to the mean within the city and, thus the degree to which activity is concentrated in particular cells. So cities 2 and 3 (ignoring city bounds) have the same degree of variation and clustering, but one that is higher than city 1 in Figure 4.

Note that the coefficient of variation has a long history, starting from Williamson (1965), for use as a measure of regional income inequality within a country. Here we are using it as a measure of economic density inequality within a city or settlement. Of course, urban economics has other measures of spatial inequality including spatial HHIs and Gini's. We focus on the coefficient of variation because it comes from a natural decomposition; and one which carries over in essence to the RPA measure.

For the RPA agglomeration measure, the decomposition is

$$RPA_j = \sum_i A_{ij} \frac{P_{ij}}{P_j} = AD_j \left(1 + \frac{Cov(A_{ij}, P_{ij})}{AD_j PD_j}\right) \tag{3}$$

where $AD_j = \frac{\sum_i^{N_j} A_{ij}}{N_j}$; $A_{ij} = \sum_{k \epsilon s} P_{kj} e^{-\alpha d_{ik}}$

In equation 3, A_{ij} is the measure over radius s of the discounted sum of neighbors ambient populations. We use an s of about 6 kms, limiting the local radius so we can distinguish later the effects of city wide versus local density. RPA_j is the weighted average of the A_{ij}, where the weights are each grid squares share of the city population. AD_j is the simple average of the A_{ij} across grid squares over the city. RPA_j can then be decomposed into the simple average, and 1 plus the covariance of A_{ij} and P_{ij}, divided by their simple averages. The latter term captures the degree to which population is allocated to grid squares with high measures of neighbors (city 2), as opposed to either being uniformly spread (city 1) or being in grid squares which are not clustered with others of high density (city 3).

In general, we will have measures of PPD_j and RPA_j at the level of a city or LD settlement. We will also have local measures, characterizing the neighborhood around which people live both for rural and urban areas, including for neighborhood i, PPD_{ij}, PD_{ij}, and A_{ij}. These we will describe in the particular contexts in which they arise. In all cases, the neighborhood of a grid cell is the square area running 5 cells to the east, west, north and south, or an area by size 11x11 grid squares (or 11x11 km which would be similar to a circle of radius 6.2).

3 How does economic density in Africa compare with the rest of the world?

Before proceeding to income and wage analyses, we see if our data support the idea that economic density in Africa is lower than in other parts of the world, despite what is presumed to be high population density in urban Africa. We interpret lower economic density as implying that, for the same overall ambient population density, there is less clustering of economic activity within African cities, so that potentially PPD_j and RPA_j, are lower, and certainly that the coefficient of variation and covariance terms in equations (2) and (3) would be lower.

We look at this for the world. To deal with issues that are pertinent outside Africa, we focused just on larger agglomerations defined in a simple fashion. Details are in Appendix

B, but effectively these areas are defined by two criteria. First, they are blobs with contiguous pixels of the density of above 1,500 per sq. km. Then, for these blobs to be in our sample, they should have at least one UN listed metropolitan area and the populations of all the listed UN metropolitan areas in the blob should sum to at least 800,000. Once we have defined these areas, we then give the agglomeration the Landscan population number obtained by summing over all grid squares in the blob. The primary issue is that, with the lower density criterion, vast swathes of seemingly rural areas in India and China are combined into, and considered, gigantic urban areas regardless of whether the areas are really urban in nature. Hence, we prefer the higher density thresholds as well as a cross check with the official UN data.For 6 African countries, we did our own checks, but doing the world in detail for smaller places and densities was beyond our scope.

Given these criteria, we establish a set of 599 cities worldwide, with 451 in the developing world. We ran regressions with dependent variables, in logs, as follows: personal population density [PPD], simple population density [PD], the coefficient of variation term in eq (2), the De La Roca-Puga agglomeration measure [RPA], the simple average of the local De La Roca-Puga measure [AD], and the covariance term in (3). For the RPA measure, we use a spatial discount rate of -0.5, as compared to De La Roca and Puga (2017) who use no discounting. Later in the paper we will analyze the optimal rate of discount for a particular and narrower context, where we find that -0.5 is close to the optimal rate for Africa.

Figure 5 shows the differences in PPD worldwide by country; where within each country we take a weighted average of each city's PPD. Blank areas are countries without cities in the data set. It is clear that African countries, in general, have very high PPD, as well as parts of South and East Asia. The question is whether that is just from high overall population density. We use simple regressions with dummies for regions of the world to answer that. Our regression results are presented in Tables 1a and 1b, where the top panel of each table gives the basic results controlling just for city ruggedness from Nunn and Puga (2012) and will represent what the raw data tell us. The bottom panel additionally controls for GDP per capita from the Penn World Tables (PWT 7.0), to see the extent to which differences in levels of development explain the patterns.

In the top panel of Table 1a, the base case is the 148 large cities in developed countries. Relative to these, Sub-Saharan African cities have higher measures across the board, including in particular the coefficient of variation and covariance terms, where they are respectively 44 and 27 per cent higher. Moreover, terms for Africa are higher than the rest of the developing world terms, including those for the coefficient of variation and covariance terms. With no separation into nocturnal and diurnal populations, we do not know if this involves greater clustering of residences or workplaces, or both; it is the ambient population as the measure of economic density. The bottom panel adds a control for ln GDPpc. This reduces the Africa terms making them smaller absolutely and relative to the rest of the developing world. Now the differentials on the coefficient of variation and covariance terms are insignificant. In summary, in the raw data Sub-Saharan African cities have higher coefficients on the coefficient of variation and covariance terms, which contradicts the presumption of the literature. Moreover, greater clustering seems to be negatively related to GDPpc, with developed countries having the lowest degree of clustering, perhaps where automobile cities like Atlanta and Houston form the stereotype.

In terms of just developing countries, Table 2, shows that relative to Asia, the outlier with lower clustering is Latin America even controlling for income. Sub-Saharan Africa, as well as North Africa and the Middle East, have similar measures of density and clustering as Asia. Overall compared to the rest of the developing world, Sub-Saharan African cities have (not controlling for GDP per capita) higher average densities of people, but no different degree of economic density as measured by PPD or RPA and no different degree of clustering. Controlling for income, Sub-Saharan African cities are similar to others in the developing world in all measures.

4 How are differences in economic density across the spatial hierarchy related to income differences?

This section first describes the data on income and wages and then the characteristics of the sample of cities and LD settlements in the covered countries. After giving the base specification, we turn to a set of results on the relationship between agglomeration measures and

income and wages, covering all areas of the country. In the next section, we will delve into looking at scale effects for cities in particular.

4.1 The data and the sample of countries and cities

We use the Living Standards Measurement Study data of the World Bank, where we have detailed geocoding of where families live for six countries; allowing us to map data to our spatial units: rural, LD settlements and cities. The LSMS surveys have detailed and consistent data at the household and individual levels on income, education, labor allocation, asset ownership, and dwelling characteristics. The data sets are the Tanzania Panel Household Survey (2008 and 2010), the Nigeria National Household Survey (2010 and 2012), the Uganda National Panel Survey (2009, 2010, 2011, and 2012), the Ethiopia Socioeconomic Survey (2011, 2013, and 2015), the Malawi Integrated Household Survey (2010 and 2013), and the Ghana Socioeconomic Panel Survey (2010 and 2013). Note that the dates of surveys in countries are so close together that they do not provide the opportunity to look at dynamics nor to identify urbanization effects off of movers.[7] These sample countries account for approximately 35% of the subcontinent's population.

Before proceeding we note how our African countries present in terms of aspects of their urban hierarchy and what the coverage of this hierarchy is by LSMS surveys. At the country level, the six countries collectively present a regular urban hierarchy. Figure 6a shows the expected (Eeckhout, 2004) log-normal distribution of all urbanized areas (cities and settlements), although there is a right tail skew. Figure 6b ranks cities from 1 to n by size with rank 1 being largest; and plots ln population against ln rank-size, so we see that rank rises (lower order) as population declines. We see that regularity holds over much of Figure 6b, governed by an approximate Pareto distribution to the right tail in Figure 6a, although the overall slope coefficient of the log-linear fit is high at -1.20, as compared to the -1 implied by the rank-size rule and the original Zipf's Law. The left tail in Figure 6a for smaller cities is an expected deviation in the right tail in Figure 6b from Zipf's Law,

[7] There is an issue of the same households appearing more than once in our data, which varies from country to country. In Table 5 below for the full sample of 44,140 households, there are 23,685 unique households, meaning that 46% of the sample involves a household that is included more than once. Clustering at the local area should remove the distortion this creates. As a robustness check, we reran Tables 5-7 with just the final year sample in the year of the LSMS for each country. Results are very similar, with similar statistical significance and coefficient magnitudes.

noting we have also bounded settlement size from below at 5,000. Note to the left in Figure 6b, that for bigger cities, the local slope coefficient would be less in absolute value than the overall -1.20, perhaps better approximating the rank-size rule.

How complete is the LSMS coverage of this hierarchy? Table 3 shows the distribution of cities with their cores and fringes broken out for our countries and for the LSMS sample. The left part of the table tells us that these countries have 167 cities (and fringes), covering 219 cores; and they have 893 LD settlements, apart from rural areas. The right part of the table shows that the LSMS data covers 115 of the 167 cities; but within these cities, only 68 fringe areas are covered. And for LD settlements only 193 of the total 893 are covered. The relatively low count of small places actually surveyed comes from the randomized sampling procedure outlined in Appendix B.

How representative is coverage by the LSMS? Figure 7 compares the size distribution of cities and LD settlements within the sample of countries versus the cities and settlements that are covered by the LSMS. The shapes of distributions of both cities and settlements for the sample are similar to those for the countries overall. The mean and median sizes for each distribution are each marked with dotted lines, with the mean being bigger than the median. For settlements, the means and medians in the LSMS sample are larger than for the country, and the same is the case for cities. This of course is consistent with Table 3.

Next, we ask about characteristics of households in the sample. Table 4 gives characteristics of the LSMS households (top panel) and working people (bottom panel) in the sample by our spatial units. Education of the household head and working-age population decline pretty sharply as we move down the spatial hierarchy. Rural areas, LD settlements, and fringes of cities are much more likely to have the household head or workers in agriculture than the core. Virtually no one anywhere is in manufacturing, the big issue for African cities (Henderson and Kriticos, 2018). Even the proportions in business services, which are potentially tradable across cities, are not that high, at 9% for cities and 2% in rural areas. Business services include the usual business service industries such as real estate and finance but add in high skill workers (like managers) in retail, as well as senior administrators and high skill workers in government. Apart from agriculture even in cities, it seems that most Africans work in low skill retail services and general labor services. However,

a key issue with the occupational data is that many people and even household heads do not report an occupation. Based on IPUMS data (Henderson and Kriticos, 2018), we believe this occurs because many of these people are farmers with agricultural land who work in other sectors as well. We note this non-reporting fraction is noticeably higher at almost 50% in rural areas.

Finally, there are the income measures. We construct measures of income for the household by adding together all income from self-employment, labor income, and capital or land income. In the surveys, respondents report income receipts of various forms, such as cash and in-kind wage payments, business incomes, remittances, incomes from the rent of property and farmland, private and government pensions, and sales revenue from agricultural produce. These receipts are also reported as taking place over a variety of time intervals, so to be consistent, we convert all income receipts to monthly intervals. Land income from crop sales or rents is generally only available at the household level, making it difficult to ascribe these income sources to any particular household member for an individual-level analysis. And the same comment applies to non-agricultural businesses owned by the household head or others in the family. For this reason, we will focus on total household income. We will also look at wage income of individuals in families which do not own agricultural land. We note that in a preliminary exercise with a smaller sample and different definitions of spatial units, we found that non-farm and farm households in cities had similar agglomeration effects to household incomes (Henderson and Kriticos, 2018). We do not explore that dimension here, especially in this sample, where the proportion of defined farm households is small.

4.2 Basic Specification and Results

All regressions have the following general specification:

$$ln(y_{ijzft}) = \alpha X_{ijzft} + \beta I_Z + \gamma_R R * S_{ijR} + \gamma_u U * S_{iju} + \gamma_c C * S_{ijc} + \delta \xi_{ft} + \epsilon_{ijzft} \tag{4}$$

- $ln(y_{ijzft})$: Income of unit i in location j of type z in country f at time t.
- X_{ijzft}: Vector of unit characteristics.

- I_Z: Vector of indicators of location type: rural[R], urbanized[U], city[C].
- S_{ijR}: Measure of rural scale within a 6km radius of unit.
- S_{iju}: Measure of overall urbanized area scale.
- S_{ijc}: Measure of city scale, as a differential from urbanized area (including settlements).
- ξ_{ft}: Vector of country-year FEs
- ϵ_{ijzft}: Error term.

We stress that what we estimate in this cross-section are correlations of income with scale measures. Any identification is from within country and year variation, and we cannot claim causal effects for two reasons. First, there is the issue of sorting by the unobserved ability across space, although that has been downplayed in the literature (Baum-Snow and Pavan, 2011). An issue is whether to control for occupation fixed effects as a way of trying to factor in ability conditional on education. While we show results with occupational fixed effects, in general, we focus on results without, because as we noted above, a large portion of our sample does not report occupation, and also because a large part of the return to being in bigger cities is the greater choice of occupations. Hence, controlling for occupation fixed effects removes this return. The second issue in terms of identification is that bigger cities may have unobserved attributes which, apart from the scale, enhance productivity, such as local public infrastructure investments. But for that, at least, the estimates will give a sense of the income pull force of cities even if scale externality effects could be overstated.

For spatial scales, we start by comparing the overall premium of being urbanized relative to rural areas, as well as the added premium of being in a city over an LD settlement. Then we turn to categorical scale measures relative to rural: which quartile of the city size distribution a household lives in, or if in a settlement, whether they are in the top or bottom 50 percentiles by settlement size.

While we start with household income, our preferred outcome measure, we will also look at wage and personal income. Table 5 presents these results. The first two columns

cover total household income for 44,140 households. Controls are listed in the table and include controls on family size and household head characteristics. The urbanized/settlement income premium is 34% and the premium for cities is 71% (0.34 + 0.37) in column 1. In column 2, settlement premiums in the two groups are similar (the average 34%); and, for cities, they have a non-monotonic pattern, ranging from 0.47 to 0.97 across the quartiles. Premiums are largest at the low and high-end sizes and are smallest for the 50-75th percentile group. This is similar to (Henderson and Kriticos, 2018) who argue that secondary cities such as in the 50-75th percentile have a role in the urban hierarchy which is limited by the lack of development of manufacturing. Below we will provide a somewhat different but not necessarily conflicting interpretation, as to why effects of population size are non-monotonic.

In columns 3 and 4 we turn to individual wage income for the 19,938 people who work over 30 hours a week for just wages, with controls listed but including hours worked. The wage premiums in cities and settlements in column 3 are similar to the household-income premiums in column 1. The quartile size ones again display a non-monotonic pattern. Finally, in columns 5 and 6 we add to wages for the full-time wage earners any non-farm business income they have and we add people who work over 30 hours a week in non-farm business activity. Results are now much weaker. We have two takeaways. First, full-time wage earners are a select group of individuals; once we add in other individuals with primarily non-wage income, urban scale premiums drop substantially. Second, looking at household income is the key; it allows returns in cities to reflect the ability of household members to work at all, to work more paid hours, and to find wage employment.

In Table 6 we turn to specifications where we experiment in each column with a different measure of economic density, proceeding from total population, to population density, personal population density and finally a De La Roca and Puga (2017) measure. Here we look just at household incomes, with controls for household characteristics but not occupation fixed effects. In Table A1 of Appendix A, we give results with no controls for household characteristics which show there are sorting effects in that scale economies generally fall when we add controls. They fall again although more modestly when we add occupation fixed effects. As noted above relative to causal effects, it is ambiguous as to

whether occupational fixed effects are appropriate. In all columns in Table 6 we allow the income intercept to vary by spatial type: rural, urbanized area and city and then we interact spatial type with a scale measure, to get continuous effects.

The first column gives classic results where scale is measured by the total population of the area, which is well defined by settlement and city boundaries. For rural areas to introduce an element of scale, we have the population of the rural area within the 11x11 squares around the households grid square. However rural scale effects are insignificant throughout. Column 1 gives two types of scale outcomes. First are marginal scale effects; here LD settlement marginal returns are surprisingly negative, and net city returns are 0.061 (0.144 - 0.083), with the latter in the range of normal estimates (Rosenthal and Strange (2008), Combes and Gobillon (2015)). Second, column 1 tells us the return to being in a city of a particular size relative to being in the rural sector. Thus, for example, relative to no scale in rural areas, cities of 5 million (15.4 in logs) pay 65% more (100*(1.09 -1.38 + 0.061*15.4)). Note that this 65% premium (in a comparatively very large city) is noticeably smaller than the 97% premium in Table 4 column 2 to being in the top quartile of cities. Below we give one explanation of why there is this difference.

Columns 2-4 of Table 5 explore different measures of density. Column 2 uses the more modern measure as in Ciccone and Hall (1996) of simple population density (and same for rural areas within the 11x11 square around a household). Now economic density elasticities become very much larger: 0.52 for LD settlements (vs -0.083 for population) and 0.52 for cities (vs. 0.061). Here density matters the same for cities and settlements.[8] In terms of a choice of economic density measure, in column 2 in Table 6, population density offers more explanatory power than population in column 1. And a horse-race between the two offers for cities and settlements a strong positive effect for density and negative or zero effect for population.[9]

In column 3 of Table 6, we explore whether using personal population density improves

[8]In comparing the elasticities in columns 1 and 2, it is important to note that a one standard deviation difference for the population is larger than for population density. For cities, one standard deviation (1.43) increase in population would increase incomes by 8.2% (vs. the elasticity of 5.75%), while a 1 standard deviation increase in population density would increase incomes by 32% (versus the elasticity of 52%). Still, the density elasticities are very high, indicating the strong pull of dense cities in potentially improving household incomes.

[9]Coefficients (s.e.s) on urban * ln pop, city* ln pop, urban *ln PD and city *ln PD are respectively -.083 (.033)**, 0.14** (0.035), 0.52** (0.16), and -0.031 (0.16).

explanatory power over the simple population density measure in column 2. For rural PPD in column 3, we have the PPD within the 11x11 square around the household. There is no improvement in Rsq in column 3 over column 2, but a horse-race weakly suggests PPD dominates PD for cities and settlements.[10] The main change with PPD is an apportioning of scale effects which now yields a smaller effect for settlements than cities, with settlement at 0.20 in column 3 versus 0.52 in column 2.[11]

In thinking about urban scale measures, columns 1 in Table 6 versus column 2 in Table 5 present two oddities noted above. First, they offer different returns of being in the biggest size cities relative to rural. Second, in Table 5, there are non-monotonic scale effects for cities of different sizes, while Table 6 column 1 suggests continuous gains to scale. These oddities are resolved by considering population density and personal population density, which provide more compelling measures of economic density than population scale. The key element is that population density measures do not rise monotonically with city population in our data. In Table 5 column 2, across the city quartiles PD and PPD are non-monotonic going from top to bottom quartile. For PD for averages, they go from 2,972, 1,711, 1,529, to 1,569, and for PPD they go from 16,980, 9,847, 10,857, to 13,324, respectively. By looking just at the population scale in explaining income, we miss the key element of density. While cities in quartiles 2 or 3 compared to the lowest quartile 4 are larger, they can have lower density, especially personal population density. The quartile specification in column 2 of Table 5 omits density considerations, and the simple population scale measure in column 1 of Table 6 does too.

Finally, we turn to the De La Roca and Puga (2017) measure in column 4 of Table 6. For RPA, we report results using a spatial discount factor of -0.7 based on the discussion in Section 5.1 below. The pattern is quite different than for PPD. There is an elasticity of 0.303 for urbanized areas, but it is the same for cities and settlements. For rural areas, we use the local measure of people market access A_{ij} in equation (3), in which j is the 11x11 square around the surveyed rural household (rather than an urban polygon).[12] Again rural areas

[10] Coefficients (s.e.s) on urban * ln PPD, city* ln PPD, urban *ln PD and city *ln PD density are respectively 0.195*** (0.0546), 0.254*** (0.0584), 0.524*** (0.155) and 0.0305 (0.157).

[11] Here the standard deviation for PD, PPD and RPA are similar for cities: 0.61, 0.71 and 0.84 respectively

[12] If we were to use A_i for other surrounding grid squares to construct a local RPA within an 11x11 square, we would have to keep expanding the area well beyond the own 11x11 square, in order to capture all relevant neighbors of those in the own square.

seem to lack density benefits. We also tried a horse-race between PPD and RPA measures, but all coefficients are insignificant with a high degree of multicollinearity.

To better explore the role of using PPD and RPA and their decomposition into mean and coefficient of variation and covariance terms, we turn to just cities. There we think spatial variation in clustering within these high-density, high population areas is likely to be more relevant, compared to settlements. And we can better distinguish local from overall city measures of clustering, since for small spatial units such as LD settlements, they are very highly correlated, based on the limited spatial extent of the unit. For example, for cities ln local PPD and ln city PPD have a simple correlation coefficient of 0.63; while for settlements it is 0.93

5 Economic density in cities

In this section, we will delve into looking at scale effects in cities in particular. In the first step, we examine whether, based on statistical criteria, there is a distance discount rate in the RPA measure which best explains income differences. Given that, we then focus on a set of issues: How important is it to distinguish coefficient of variation and covariance terms as well as a spatial Gini from simple density terms in explaining income differences? In cities, do local measures of density additionally explain income differences across households? Do people in the fringe versus core areas of cities experience different agglomeration effects?

5.1 Constructing de la Roca-Puga measures

We take the specification in column 3 of Table 5 and drop all rural and city terms, so for our sample of cities, we are left with just the controls and country-year fixed effects and the ln RPA term. We start with spatial discount rates of -0.1 and raise that in absolute value in increments of 0.1 to -1.5. -1.5 is an extremely high discount: at 1km distance, neighbors have a weight of 0.22; at 2km it is already only 0.05, and by 5km their weight is effectively 0. For each discount rate, we record the F-value of adding the ln RPA term; these values are shown in Figure 8. The peak is at -0.7, so that the improvement in explaining income differences across cities is maximized at the -0.7 discount rate. We do note values from -0.5

to -0.9 yield similar Fs. Throughout for the Africa work we use the discount rate of -0.7 in all cases for any type of RPA measure. For the world, we used the more conservative rate with lower discounting of -0.5. The Africa results for -0.5 and -0.7 are very similar across the board.

We could have done a differently specified F-test where we decompose the RPA measure into its ln AD and ln (1 + covariance term) in equation 3. The results for that give a corner maximum F at a discount rate of -1.5. Hence, it is only at very high discount rates that the covariance term becomes significant. However, such a high discount rate says neighbors do not matter and effectively reduces RPA to something close to PPD. We have two takeaways. First, the indication is that neighbors, in general, are not so important. Even at a spatial discount of -0.7, at 2km, neighbors only get a weight of 0.25. Second and related, the test suggests that perhaps we may want to focus on PD or PPD, rather than RPA.

5.2 Economic density results for cities

We have three sets of results: the first involves second moment and other dispersion measures; the second concerns local density effects, controlling for overall city density; and finally, the third concerns whether effects differ by location in the city.

5.2.1 Second moment measures

Does the degree of clustering in cities matter, at least as we currently can measure it in this developing country context? In Table 7 we look at people living just in cities and at their returns to clustering. In columns 1 and 2 we first repeat what in essence are the column 2 and 3 regressions in Table 6 for just cities with the measure of ln PD and ln PPD. Compared to Table 6 we get very similar overall elasticities of 0.59 and 0.43 respectively, with ln PD in column 1 explaining more of the variation in the data. In column 3 in Table 7, we decompose ln PPD into the ln PD and ln (1 + coefficient of variation term) in equation 2. The covariance term is small and insignificant and thus does not add to explanatory power relative to just using ln PD in column 1. In columns 4-6 we repeat the same exercises with the RPA measure getting the same type of results. We also conducted a series of horseraces

which suggest the ln PD term dominates ln AD, ln RPA and ln PPD.[13] In short the measures of the differential degree of clustering within cities do not seem to add to the analysis.

There are two issues with drawing the conclusion that we should focus on PD rather than PPD or RPA measures for cities. First concerns measurement error. Use of Landscan data measures within-city clustering and inequality with error. While Landscan may do a better job than other currently available data sets, the assignment of people to work and residential locations is surely done with considerable error, which would bias the coefficients downward. The second issue concerns our measure of clustering. While our measure of clustering fits into a decomposition and is a standard measure in looking at spatial inequality, there are other standard measures. We look at one of these. In a context with so many grid cells in each urban area but a varying number, we preferred the spatial Gini to use incomparison over HHI based ones or a Theil index. We calculated the spatial Gini of cell concentration of economic activity within each city.[14] In column 7 of Table 7, the Gini has a completely insignificant coefficient. Of course, as for the coefficient of variation and covariance terms, the small size of the coefficient relative to the standard error could be explained by measurement error.

The conclusion for this sample and data is that a simple population density measure works as well as more nuanced measures, in attempting to capture economic density of cities in explaining income differences.

5.2.2 Does neighborhood density matter?

A key issue as noted in the introduction is that studies suggest that within cities there are local scale externalities which decay sharply with distance so that from that perspective firms in one neighborhood do not interact with firms in another (Arzaghi and Henderson,

[13] In terms of horseraces for ln PD vs ln AD the ln PD term is positive at 0.74 (but with an s.e. of 0.44), while the ln AD term is small (-0.145) and insignificant. For ln PD vs ln PPD, as column 3 already tells us, the ln PD coefficient is large (0.51) and significant, while that for ln PPD is small (0.090) and insignificant. Finally for ln PD, against ln RPA, ln PD has a coefficient of 0.489 (but with an s.e. of 0.283) while ln RPA has a small (0.077) insignificant coefficient.

[14] We do this by ordering each cell by its density and noting the cumulative share of the population in each cell. The cumulative distribution of the population ordered by density represents the Lorenz curve. We then sum up the area under the Lorenz curve (by adding up the "height" and the "width" of each bar of the cumulative population histogram, where the "height" is the cumulative population and the "width" is 1/(number of total cells in the city)). We call this integral I, and, according to the Gini formula, calculate Gini=1-2I for each city.

2008). That begs the question then of why there are two neighborhoods of seemingly non-interacting firms found in one city. The answer must be that firms benefit more generally from the urban scale.

In Table 8 we address this issue for household income data. All columns control for citywide ln PD. Local density and overall measures are correlated, so if the local density measures matter that will reduce the elasticity (0.59 in col 1) of ln city PD. This is the case except in column 4 where the local density measure is insignificant. While the coefficients on ln city PD decline when we include key local measures, they are still very large, indicating large marginal returns to overall city density. Columns 2-5 each experiment with a different measure of local economic density. Although we have argued that a simple PD measure works as well or better than anything else for a citywide measure, that does not mean at the local level it is the best measure. In column 2 the local measure is ln (local PD); in 3 it is ln (local PPD); in column 4 it is ln (own cell population), and in 5 it is the simple de la Roca-Puga measure (A in eq. 3). The ln (own cell population) has an insignificant effect. In all other columns, local density measures have strong significant effects, with little to choose between them in terms of magnitudes. Magnitudes are all about 0.14, much smaller than the elasticity of about 0.45 for overall city density. R-squared's are very similar across columns 2, 3 and 5, although the column with ln local PPD does have the highest R-squared. Horseraces suggest ln local PPD with a coefficient (s.e.) of 0.135 (0.071) dominates ln local PD (0.038 (0.068)) and ln local PPD with a coefficient of 0.141 (0.058) dominates ln local A (0.030 (0.053)).

In summary, for a household, the overall density measure for the city has strong effects on incomes, but also the local area around the household has strong effects as well. Households benefit from both dense overall cities and dense 11 x 11 km square neighborhoods.

5.2.3 The core versus fringe of cities

Finally, in this section, we examine the role of the core versus fringe of cities. Do people benefit from overall city PD or really just the PD of the core? Do people living in the fringe benefit differently from overall or local density effects? We look at this in Table 9. Columns 1 and 2 are very important. Column 1 shows that, while people in the fringe of

cities earn less controlling for household characteristics, overall city scale economies are very important for them. While people in the core have a density elasticity of 0.45 those in the fringe have a big extra kick and total elasticity of 0.75. Column 2 tells us that it is more overall city density that matters not density in the core. Column 3 then adds a local density measure. While people in the fringe still get an extra kick relative to those in the core from overall city density, we can not establish that local effects for those in the fringe are greater than those in the core. Column 4 reruns column 3 showing a similar core versus fringe differential response to overall city density, but it imposes the same effects in the fringe and core for local density.

The puzzle is the bigger role of city density externalities for fringe than core city residents. We thought this was because of sorting based on migration status, where migrants lack information and might benefit more from say information spillovers. Hence, we expected that the fringe would have a greater proportion of migrants (defined by having moved within the last 5 years) than the core within the LSMS sample. However, there is only a tiny difference: 16% vs 15%. Moreover, coefficients on migration status interacted with scale effects are insignificant and have no impact on the fringe results. So that leaves a puzzle. It certainly says that including the fringe of a city is important, but we do not have a ready answer for why economic density effects may be greater for people in the fringe. There certainly must be sorting on some dimension, where people in the fringe earn less for the same observables, but somehow this disadvantaged population benefits more from greater city density, even though by construction they live in less dense neighborhoods.[15]

6 Looking at firm productivity within a city: Kampala

This section will highlight a critical aspect of the work in this paper. What we have investigated is the return to density in improving household incomes, including benefits from local neighborhood density within cities. Households are the ultimate beneficiaries of economic density in terms of wages and business incomes. In the literature as noted in the introduction there are a set of papers looking at income returns from agglomeration. How-

[15]These results apply to a sample excluding Nigeria where we do not know migration status. Results in the version of Table 9 without Nigeria are similar to the current Table 9.

ever there is also a literature looking at firm productivity. The impacts of economic density may be different for firms than for households. Obviously firm productivity directly affects incomes earned, but incomes also reflect labor force participation and matching opportunities for the household, as well as incomes from the informal sector which are not well represented in censuses or typical surveys.

While we found strong neighborhood externalities for different density measures in Table 8, we find no such effects for these types of measures for firms. For firms being in a neighborhood with more people, total employment or Landscan ambient population does nothing to firm productivity measured by value added per worker. The only impact on productivity is from neighborhood localization externalities – having more employment or more firms within a defined neighborhood near a firm within the own industry. And, as we will see even those effects are nuanced, applying to particular sectors and trading off competition versus spillover effects.

6.1 Kampala data

To study firm productivity at the neighborhood level in Kampala, we use data from the Uganda Business Inquiry (UBI) survey conducted in 2002. The UBI is an economic survey which made use of the official Census of Business Establishments (COBE) of 2002 as its sampling frame. The principal objective of the survey is to provide the necessary information and data to measure the contribution of each industry sector to the growth of the economy. So the survey covers a large range of economic variables, including value added and business assets. Coverage is comprehensive, with information on all sectors of the economy - including the informal sector[16] - and coverage of all the officially recognized districts in Uganda. The sector definitions are in line with the International Standard Industrial Classification (ISIC), Revision 4, and cover 15 1-digit sectors.[17] A stratified two-stage sample design was used to select the businesses for the UBI, which focused on getting

[16]The informal status of a business is recorded using a question in the survey on whether a business pays any taxes as its determinant; this is in line with other surveys conducted in Uganda such as the UNHS which were more specifically directed to studying informality in the economy.

[17]These are agriculture & fishing, mining & quarrying, manufacturing, construction, utilities, trade, transport & storage, accommodation & food services, information & communication, finance & insurance, real estate & business services, education, health & social work, recreation and personal services. Note that the survey excludes information on the following ISIC sections: Section O, i.e., public administration and defense; compulsory social security, and Section U, i.e., activities of extraterritorial organizations and bodies.

all bigger establishments with over 50 employees and then sampling at the lower end.[18]

Although the UBI covers the entire country, due to confidentiality issues and limits on the capacity to share this data outside of the Uganda Bureau of Statistics (UBOS), our data have coverage only of Kampala and furthermore we are restricted by the abridged version of the data and statistics that have been provided to us. Our data cover the Greater Kampala metropolitan area for a sample of 2,342 firms, each with information on their GPS location, total number of employees, value added per worker, and industry classification. Value added is calculated as the net output of a sector after adding up all outputs and subtracting out all intermediate inputs but labor.

6.2 Matching the data to density measures

We measure the effects of density on firm productivity by matching the firm coordinates and data from the UBI survey with our population, employment, and Landscan ambient population data, used elsewhere in this paper. As detailed in section 1, our data on population are from the 2002 census and are at the level of 174 parishes within the Greater Kampala administrative area. We assign that data to the 1km grid square level, through a weighted sum to the survey area numbers, so as to be consistent with the spatial resolution of Landscan. Our employment data are from the 2002 Census of Business Establishments [COBE] and provides the GPS location of supposedly the universe of firms and employment in Greater Kampala.[19]

The data allow us to study density effects at two spatial scales. First is the coarser one with 1km grid squares upon which population and Landscan data are defined. For this each firm point in the UBI dataset is overlaid by the 1km grid square level data on population, employment, and Landscan. Each firm is then assigned own-square measure

[18] Business establishments were first stratified by industry sector and within each industry sector, business establishments were further stratified by employment size as per the following categories of employment size 1, 2-4, 5-9, 10-19, 20-49, 50-99, 100-499 and 500 or more employees; and further by turnover; thus less than 5 million shillings, between 5 and 10 million shillings and more than 10 million shillings. Given the significant contribution of the larger establishments, from previous surveys, to the value of production; all the establishments with 50 or more employees were supposed to be sampled with certainty, (a probability of 1), while those employing fewer than 50 employees were subjected to probabilistic sampling.

[19] An issue is that if we look in the 2002 census data, matching it alongside the firm locations from our 2002 survey, we see that for over half of our firms we cannot find other firms from the census that are in the same industry and within 150 meters of the survey firms, indicating either that there is a lot of firm turnover or issues in recording locations.

based on the grid square in which it is located, as well as local neighborhood measures from the grid squares surrounding each firm's own-cell. Specifically, we define three rings around the own-cell to capture local neighborhood effects and the extent of spatial decay; as illustrated in Figure B4a of Appendix B, ring 1 captures the 8 cells in the immediate queen neighborhood of the reference cell, expanding incrementally by one cell in every direction. Ring 2 captures the 16 cells around ring 1, and finally ring 3 expands to the 24 cells around ring 2. For each ring, we calculate the mean density of 2002 population, 2002 employment and 2012 Landscan [20] For the 2002 COBE we also start with that scale for total employment and own industry employment. However a second spatial scale is possible for employment, since we know point locations. For each survey firm we can form precise circular rings of all firms within any distance. We then count employment and plants within each of those rings. We experimented with fine rings in 500 meter increments; but, for the own industry, firm counts in those distances were sparse so we settled on 1km increments with 4 rings: 0-1km , 1-2km, 2-3km and 3-4km.

Table 10 pools all types of industries and uses the coarser spatial scale. It presents the results of regressions of the log of value added per worker for each firm in our sample, on measures of density in the firm's own-cell and its respective rings. For each regression, we include industry fixed effects and controls for distance to the city center, Lake Victoria, and the Kampala Northern Bypass. Column 1 presents results using population as the measure of density and shows that there are no significant effects on firm value added, with no hint at a strong explicable pattern. Similarly, in column 2, we detect no significant effects on value added from Landscan ambient population. Our results for employment are more interesting. First, they suggest that it is employment density that matters for firm productivity, not the resident population or a composite of the resident population. In column 3, there appears to be a negative effect of greater employment density in the own-cell, with positive spillover effects occurring from the second ring in the radius around 2km away from the reference cells. Experimentation suggested employment effects were driven by own industry effects. In column 4 on own-employment density, we also see the

[20] Obviously the Landscan data are for much later. We do note the simple correlation cofficients on own indsutry employment in Table 11 right panel, between 2002 and 2012 census measures are all over 0.97 for the 4 rings.

contrasting negative and positive effects from the own-cell and second ring respectively. Table 11 below, with more precise ring employment data, will also suggest that there are negative competition effects of employment in the own cell and positive spillovers beyond that.

We then looked at these effects by sub-industry. In general we found no effects of population or Landscan measures on any outcomes for any sub-industry, including no effects of nearby population on retail trade or personal services. The same comment applies to employment measures with one set of exceptions. These are industries which can be characterized as producing traded goods: manufacturing and business services. We expect that such firms are likely to benefit differently from density compared to other industries. Manufacturing, in particular, as well as business services are archetypal industries that are not only exportable in scope, but also input and/or employee intensive. Hence, such firms may not need to be close to local population and consumers, but they need good connectivity to transport infrastructure and employment. Moreover, such firms are likely to benefit more positively from knowledge spillovers and labor sharing, meaning there is a greater advantage of these firms to cluster together.

In Table 11, we focus on this sub-sample of firms. To maintain sample size, we pooled these two sets of industries. Table 11 in the left panel shows that manufacturing and business services firms in our sample see no benefit from greater population or Landscan density. However, they experience greater positive effects from employment and own-employment density in the second and third rings, which compete against negative effects from greater local density in the own-cell. We think the own ring is dominated by negative competition effects whether in input markets (poaching of employees and suppliers) or output markets (shopping externalities with loss of customers to competitors and negative price effects). Those effects appear to apply to overall employment even more than own industry. Beyond the first two rings, positive spillovers dominate and have very strong elasticities, especially at ring 2. In the right panel of Table 11 we explore this more with detailed own industry data, looking at precise rings based on point locations and trying both own industry employment and own industry count of firms as the scale measure. In both cases, we find significant negative, competition effects in the own ring, and positive

and significant spillovers effects in the third ring at 2-3km. In the second ring at 1-2km the two forces seem to offset each other. And by ring 4 beyond 3km any effects have dissipated. These results are similar to those for advertising in Manhattan in Arzaghi and Henderson (2008), except the succession as we move away from the own firm of competition, spillover, and dissipation occurs at much shorter distances than in Kamapala.

7 Conclusions

This paper evaluates the use of different measures of economic density in assessing urban agglomeration effects. These density measures are based on Landscan data. While Landscan measures ambient population at the grid square level with error, a ground-truthing exercise for Nairobi and Kampala suggests that Landscan does a better job of capturing within city clustering than other measures derived from smearing population into grid squares based on ground cover such as Landsat type data. The empirical work focuses on Africa, where there is a perception that African cities have low economic density due to a lack of clustering of economic activity. We find that Africa as a region has economic density and cluster measures similar to other developing countries in Asia and North Africa and higher than those found in the developed world or Latin America.

To assess economic density measures, we examine how well they explain household income differences across cities and neighborhoods. We have simple scale and density measures and more nuanced ones which capture in second moments the extent of clustering within cities. Noting that the extent of clustering is measured with error in Landscan resulting in attenuation bias, the evidence we have suggests that a simple density measure explains income differences across cities as well as or even better than more nuanced measures attempting to capture within city differences in the extent and nature of clustering. However, simple city scale measures such as total population are inferior to density measures and to some degree misleading. On the big picture side, there are very large household income premiums from being in bigger and particularly denser cities over rural areas in Africa, indicating migration pull forces remain very strong. Second, the marginal effects of increases in density on household income are very large, with density elasticities close of 0.6.

Besides overall city density measures, we look at density in the neighborhood around a household within a city. In addition to strong city-level density effects we find strong neighborhood effects looking at neighborhoods of about 6km radius. The elasticity of overall city density is 0.43 and for local density it is 0.14. Both overall city density and density of the own neighborhood matter.

References

Arzaghi, M. and J. V. Henderson (2008). Networking off madison avenue. *The Review of Economic Studies 75*(4), 1011–1038.

Barrios, S., L. Bertinelli, and E. Strobl (2006). Climatic change and rural–urban migration: The case of sub-saharan africa. *Journal of Urban Economics 60*(3), 357–371.

Baum-Snow, N. and R. Pavan (2011). Understanding the city size wage gap. *The Review of Economic Studies 79*(1), 88–127.

Brückner, M. (2012). Economic growth, size of the agricultural sector, and urbanization in africa. *Journal of Urban Economics 71*(1), 26–36.

Bustos, P., B. Caprettini, and J. Ponticelli (2016). Agricultural productivity and structural transformation: Evidence from brazil. *American Economic Review 106*(6), 1320–65.

Ciccone, A. and R. E. Hall (1996). Productivity and the density of economic activity. *The American Economic Review 86*(1), 54–70.

Collier, P. and P. Jones (2016). Transforming dar es salaam into a city that work. *Tanzania: The Path to Prosperity*, 86.

Collier, P., P. Jones, and D. Spijkerman (2018). Cities as engines of growth: Evidence from a new global sample of cities. *Unpublished*.

Combes, P.-P. and L. Gobillon (2015). The empirics of agglomeration economies. In *Handbook of regional and urban economics*, Volume 5, pp. 247–348. Elsevier.

De La Roca, J. and D. Puga (2017). Learning by working in big cities. *The Review of Economic Studies 84*(1), 106–142.

Desmet, K., J. Gomes, and I. Ortuño-Ortín (2018). The geography of linguistic diversity and the provision of public goods. Technical report, National Bureau of Economic Research.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review 94*(5), 1429–1451.

Fay, M. and C. Opal (1999). *Urbanization without growth: a not-so-uncommon phenomenon*. The World Bank.

Fujita, M. and H. Ogawa (1982). Multiple equilibria and structural transition of non-monocentric urban configurations. *Regional Science and Urban Economics 12*(2), 161–196.

Galor, O. and A. Mountford (2008). Trading population for productivity: theory and evidence. *The Review of Economic Studies 75*(4), 1143–1179.

Glaeser, E. L. and D. C. Mare (2001). Cities and skills. *Journal of Labor Economics 19*(2), 316–342.

Gollin, D., R. Jedwab, and D. Vollrath (2016). Urbanization with and without industrialization. *Journal of Economic Growth 21*(1), 35–70.

Gollin, D., S. L. Parente, and R. Rogerson (2007). The food problem and the evolution of international income levels. *Journal of Monetary Economics 54*(4), 1230–1255.

Hansen, G. D. and E. C. Prescott (2002). Malthus to solow. *American Economic Review 92*(4), 1205–1217.

Henderson, J. V. and S. Kriticos (2018). The development of the african system of cities. *Annual Review of Economics 10*.

Henderson, J. V., T. Regan, and A. Venables (2018). Building the city: Urban transition and institutional frictions. *Unpublished*.

Henderson, J. V., A. Storeygard, and U. Deichmann (2017). Has climate change driven urbanization in africa? *Journal of Development Economics 124*, 60–82.

Lall, S. V., J. V. Henderson, and A. J. Venables (2017). *Africa's cities: Opening doors to the world*. World Bank Publications.

LandScan (2012). *Oak Ridge National Laboratory https://landscan.ornl.gov/*.

Lewis, W. A. (1954). Economic development with unlimited supplies of labour. *The Manchester School 22*(2), 139–191.

Matsuyama, K. (1992). Agricultural productivity, comparative advantage, and economic growth. *Journal of Economic Theory 58*(2), 317–334.

Nunn, N. and D. Puga (2012). Ruggedness: The blessing of bad geography in africa. *Review of Economics and Statistics 94*(1), 20–36.

OECD (2012). *Redefining "Urban": A New Way to Measure Metropolitan Areas*.

Rosenthal, S. S. and W. C. Strange (2004). Evidence on the nature and sources of agglomeration economies. In *Handbook of Regional and Urban Economics*, Volume 4, pp. 2119–2171. Elsevier.

Rosenthal, S. S. and W. C. Strange (2008). The attenuation of human capital spillovers. *Journal of Urban Economics 64*(2), 373–389.

Schultz, T. W. et al. (1968). Economic growth and agriculture. *Economic Growth and Agriculture*.

Small, C. and J. Cohen (2004). Continental physiography, climate, and the global distribution of human population. *Current Anthropology 45*(2), 269–277.

Venables, A. J. (2018). Urbanisation in developing economies: Building cities that work. *REGION 5*(1), 91–100.

Williamson, J. G. (1965). Regional inequality and the process of national development: a description of the patterns. *Economic Development and Cultural Change 13*(4, Part 2), 1–84.

Table 1: Density Measures for Africa vs Other World Regions

	(1) PPD	(2) Pop. Density	(3) 1 + CV Term	(4) RPA (5km, e=-0.5)	(5) RPA NW (5km, e=-0.5)	(6) 1 + Cov. Term
Sub-Saharan Africa	1.101***	0.662***	0.440***	0.930***	0.665***	0.265***
	(0.139)	(0.101)	(0.0778)	(0.128)	(0.0908)	(0.0557)
Rest of DevelopingWorld	0.903***	0.564***	0.339***	0.792***	0.573***	0.219***
	(0.110)	(0.0885)	(0.0457)	(0.111)	(0.0724)	(0.0425)
Ln(Ruggedness)	0.0938***	0.0535***	0.0403**	0.0574***	0.0475***	0.00988
	(0.0244)	(0.0143)	(0.0197)	(0.0178)	(0.0141)	(0.00841)
Ln(Pop. Landscan)	0.179***	0.164***	0.0152	0.291***	0.198***	0.0934***
	(0.0162)	(0.0129)	(0.0171)	(0.0146)	(0.0154)	(0.0102)
Observations	602	602	602	602	602	602
R-squared	0.714	0.577	0.629	0.707	0.661	0.548
Panel B: GDP Control						
Sub-Saharan Africa	0.541***	0.411***	0.130	0.441***	0.347***	0.0944
	(0.183)	(0.130)	(0.149)	(0.149)	(0.119)	(0.0797)
Rest of DevelopingWorld	0.618***	0.436***	0.182**	0.543***	0.410***	0.133***
	(0.106)	(0.0868)	(0.0832)	(0.0960)	(0.0712)	(0.0458)
Ln(Ruggedness)	0.0991***	0.0565***	0.0426**	0.0625***	0.0513***	0.0111
	(0.0236)	(0.0151)	(0.0182)	(0.0179)	(0.0148)	(0.00780)
Ln(GDP pc)	-0.180***	-0.0794**	-0.100**	-0.156***	-0.101***	-0.0552**
	(0.0495)	(0.0346)	(0.0449)	(0.0386)	(0.0325)	(0.0216)
Ln(Pop. Landscan)	0.168***	0.158***	0.00974	0.282***	0.191***	0.0903***
	(0.0158)	(0.0123)	(0.0165)	(0.0150)	(0.0150)	(0.0104)
Observations	599	599	599	599	599	599
R-squared	0.735	0.588	0.646	0.724	0.675	0.567

Notes: Table reports results from OLS regressions; errors are clustered by country. Asterisks indicate $p < 0.01(***)$, $p < 0.05(**)$, and $p < 0.1(*)$.

Table 2: Density Measures: Developing Countries Only

	(1) PPD	(2) Pop. Density	(3) 1 + CV Term	(4) RPA (5km, e=-0.5)	(5) RPA NW (5km, e=-0.5)	(6) 1 + Cov. Term
Sub-Saharan Africa	0.0854	0.137***	-0.0520	0.118	0.141***	-0.0224
	(0.104)	(0.0522)	(0.0764)	(0.0821)	(0.0454)	(0.0564)
Latin America	-0.211**	0.0638	-0.274***	-0.0693	0.0771	-0.146***
	(0.0918)	(0.0505)	(0.0834)	(0.0697)	(0.0584)	(0.0496)
North Africa	0.214	0.127	0.0873	0.162	0.0150	0.147
	(0.220)	(0.175)	(0.115)	(0.225)	(0.189)	(0.0907)
Ln(Ruggedness)	0.0796***	0.0409**	0.0387*	0.0534**	0.0440***	0.00949
	(0.0272)	(0.0163)	(0.0218)	(0.0216)	(0.0162)	(0.0112)
Ln Pop Landscan, thsnds	0.152***	0.155***	-0.00331	0.225***	0.171***	0.0542***
	(0.0240)	(0.0173)	(0.0250)	(0.0179)	(0.0184)	(0.0162)
Observations	454	454	454	454	454	454
Panel B: GDP Control						
Sub-Saharan Africa	-0.0444	0.0712	-0.116	-0.0220	0.0406	-0.0626
	(0.132)	(0.0645)	(0.112)	(0.0976)	(0.0558)	(0.0800)
Latin America	-0.118	0.114*	-0.232**	0.0325	0.152**	-0.120**
	(0.103)	(0.0614)	(0.0901)	(0.0832)	(0.0645)	(0.0489)
North Africa	0.178	0.108	0.0696	0.123	-0.0132	0.136
	(0.202)	(0.165)	(0.112)	(0.205)	(0.174)	(0.0895)
Ln(Ruggedness)	0.0843***	0.0442***	0.0401*	0.0589***	0.0484***	0.0104
	(0.0269)	(0.0167)	(0.0207)	(0.0220)	(0.0168)	(0.0110)
Ln GDP per capita	-0.115**	-0.0595	-0.0555	-0.125**	-0.0897**	-0.0350
	(0.0551)	(0.0407)	(0.0572)	(0.0482)	(0.0380)	(0.0340)
Ln Pop Landscan, thsnds	0.151***	0.154***	-0.00345	0.224***	0.170***	0.0541***
	(0.0238)	(0.0169)	(0.0249)	(0.0168)	(0.0178)	(0.0159)
Observations	451	451	451	451	451	451

Notes: Table reports results from OLS regressions; errors are clustered by country. Asterisks indicate $p < 0.01(***)$, $p < 0.05(**)$, and $p < 0.1(*)$.

Table 3: Counts of Urbanized Areas in Our Countries and Sample

	All urban areas in our countries				*Urban Areas where surveyed units live*			
	Cities	**Cores**	**Fringes**	**LD**	**Cities**	**Cores**	**Fringes**	**LD**
Ethiopia	24	33	24	223	16	19	5	33
Ghana	14	15	14	53	10	8	6	15
Malawi	5	4	5	49	3	3	4	30
Nigeria	92	128	92	389	58	62	31	44
Tanzania	23	24	23	50	14	11	8	15
Uganda	9	15	9	129	14	13	14	56
Total	167	219	167	893	115	116	68	193

Table 4: Household and Person Characteristics by Location

	City	Core	Fringe	Settlement	Rural
Panel A					
Household head in Agriculture	0.124	0.0629	0.273	0.261	0.297
Household head in Business Services	0.0885	0.108	0.0418	0.0510	0.0238
Household head in Manufacturing	0.0271	0.0297	0.0205	0.0220	0.00877
Household head in Not-Recorded	0.362	0.375	0.329	0.397	0.561
Household head with >Primary Education	0.441	0.503	0.290	0.307	0.173
Panel B					
Working age pop. in Agriculture	0.0925	0.0417	0.218	0.250	0.291
Working age pop. in Business Services	0.0745	0.0900	0.0361	0.0401	0.0187
Working age pop. in Manufacturing	0.0287	0.0312	0.0224	0.0319	0.0172
Working age pop. in Not-Recorded	0.409	0.412	0.401	0.419	0.559
Working age pop. with >Primary Education	0.495	0.546	0.371	0.319	0.172
No. Urban Areas	115	116	68	193	-

Table 5: Estimation of Household and Individual Income Premiums

	Household Level		**Individual Level**			
	Total Income		*Wage Income*		*Total Income*	
	(1)	(2)	(3)	(4)	(5)	(6)
Urban	0.338***		0.374***		0.213***	
	(0.0313)		(0.0474)		(0.0346)	
City	0.373***		0.350***		0.173***	
	(0.0339)		(0.0443)		(0.0336)	
Cities Above 75th percentile		0.972***		0.766***		0.444***
		(0.0299)		(0.0309)		(0.0256)
Cities between 50th-75th percentile		0.468***		0.591***		0.290***
		(0.0454)		(0.0422)		(0.0391)
Cities between 25th-50th percentile		0.481***		0.868***		0.386***
		(0.0398)		(0.0460)		(0.0361)
Cities between below 25th percentile		0.656***		0.579***		0.313***
		(0.0354)		(0.0381)		(0.0319)
LD Settlements Above 50th percentile		0.308***		0.413***		0.228***
		(0.0427)		(0.0629)		(0.0457)
LD Settlements Below 50th percentile		0.375***		0.330***		0.198***
		(0.0425)		(0.0631)		(0.0466)
Observations	44140	44140	19938	19938	22842	22842
R^2	0.331	0.334	0.158	0.160	0.224	0.225
Country-Year FE	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓

Notes: Table reports results from OLS regressions; the dependent variable in columns (1)-(2) and (5)-(6) is the natural logarithm of total net income from all available sources. The dependent variable in columns (3)-(4) is the natural logarithm of wage income. Controls at the household level are the education (recorded or not), level of education (if recorded), age, age squared and gender of the household head; household size; household size squared; whether the household owns land and, if so, the natural logarithm of the size of land holdings in hectares. Controls at the individual level are education (recorded or not); level of education (if recorded); age; age squared; gender; hours worked (recorded or not); and number of hours worked (if recorded). The sample in columns (3)-(6) is limited to individuals working more than 30 hours a week. Each column includes country-year fixed effects. Robust standard errors are presented in parentheses. Asterisks indicate $p < 0.01(***)$, $p < 0.05(**)$, and $p < 0.1(*)$. *Source: World Bank Living Standard Measurement Surveys.*

Table 6: Estimation of Economic Density Effects on Household Income Premiums

	Ln(Population)	**Ln(Pop. Density)**	**Ln(PPD)**	**Ln(RPA)**
	(1)	(2)	(3)	(4)
Rural X Scale	-0.0148	-0.0213*	0.0157*	-0.00853
	(0.0108)	(0.0112)	(0.00825)	(0.0104)
Urban X Scale	-0.0827**	0.524***	0.195***	0.303***
	(0.0327)	(0.155)	(0.0546)	(0.0797)
City X Scale	0.144***	0.0305	0.254***	0.0890
	(0.0346)	(0.157)	(0.0584)	(0.0815)
Urban	1.090***	-3.198***	-1.134**	-2.535***
	(0.368)	(1.025)	(0.445)	(0.748)
City	-1.376***	-0.306	-2.198***	-1.048
	(0.388)	(1.038)	(0.484)	(0.767)
Observations	44140	44140	44118	44140
R-squared	0.332	0.338	0.337	0.337
Country-Year FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of total net household income from all available sources. Controls are the education (recorded or not), level of education (if recorded), age, age squared and gender of the household head; household size; household size squared; whether the household owns land and, if so, the natural logarithm of the size of land holdings in hectares. Each column includes country-year fixed effects. Standard errors are clustered at the individual urbanized area level. Asterisks indicate $p < 0.01(***)$, $p < 0.05(**)$, and $p < 0.1(*)$. *Source: World Bank Living Standard Measurement Surveys.*

Table 7: Estimation of Density and Clustering Effects on Household Income Premiums

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Ln(Pop. Density)	0.596***		0.602***				
	(0.0782)		(0.0797)				
Ln(PPD)		0.436***					
		(0.0836)					
Ln(1 + CV Term)			0.0802				
			(0.134)				
Ln(AD)				0.574***		0.553***	
				(0.0745)		(0.0972)	
Ln(RPA)					0.406***		
					(0.0608)		
Ln(1 + Cov Term)						0.0705	
						(0.257)	
Gini							0.209
							(0.694)
Observations	11227	11227	11227	11227	11227	11227	11227
R^2	0.275	0.269	0.276	0.275	0.273	0.275	0.253
Country-Year FE	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of total net household income from all available sources. Controls are the education (recorded or not), level of education (if recorded), age, age squared and gender of the household head; household size; household size squared; whether the household owns land and, if so, the natural logarithm of the size of land holdings in hectares. Each column includes country-year fixed effects. Standard errors are clustered at the individual urbanized area level. Asterisks indicate $p < 0.01(***)$, $p < 0.05(**)$, and $p < 0.1(*)$. *Source: Author's dataset*

Table 8: Estimation of Average and Local Density Effects on Household Income Premiums

	(1)	(2)	(3)	(4)	(5)
Ln(City PD)	0.587***	0.431***	0.472***	0.566***	0.462***
	(0.0775)	(0.0946)	(0.0849)	(0.0792)	(0.0886)
Ln(Local PD)		0.139***			
		(0.0451)			
Ln(Local PPD)			0.161***		
			(0.0461)		
Ln(Own-Cell Population)				0.0210	
				(0.0159)	
Ln(Local A)					0.121***
					(0.0427)
Observations	11493	11493	11493	11493	11493
R^2	0.275	0.277	0.278	0.275	0.277
Country-Year FE	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of total net household income from all available sources. Controls are the education (recorded or not), level of education (if recorded), age, age squared and gender of the household head; household size; household size squared; whether the household owns land and, if so, the natural logarithm of the size of land holdings in hectares. Each column includes country-year fixed effects. Standard errors are clustered at the individual urbanized area level. Asterisks indicate $p < 0.01(***)$, $p < 0.05(**)$, and $p < 0.1(*)$. *Source: Author's dataset*

Table 9: Estimation of Core vs Fringe Density Effects on Household Income Premiums

	(1)	(2)	(3)	(4)
Ln(City PD)	0.458***		0.344***	0.331***
	(0.0759)		(0.106)	(0.0970)
Ln(Core PD)		0.528***		
		(0.0886)		
Ln(Local PD)			0.108*	0.120**
			(0.0652)	(0.0547)
Fringe X Ln(City PD)	0.278**		0.371***	0.395***
	(0.112)		(0.140)	(0.128)
Fringe X Ln(Core PD)		-0.0772		
		(0.134)		
Fringe X Ln(Local PD)			0.0787	
			(0.140)	
Fringe	-2.332***	0.283	-3.405***	-3.041***
	(0.839)	(1.086)	(1.071)	(0.926)
Observations	11232	11232	11232	11232
R^2	0.280	0.275	0.281	0.281
Country-Year FE	✓	✓	✓	✓
Controls	✓	✓	✓	✓

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of total net household income from all available sources. Controls are the education (recorded or not), level of education (if recorded), age, age squared and gender of the household head; household size; household size squared; whether the household owns land and, if so, the natural logarithm of the size of land holdings in hectares. Each column includes country-year fixed effects. Standard errors are clustered at the individual urbanized area level. Asterisks indicate $p < 0.01(***)$, $p < 0.05(**)$, and $p < 0.1(*)$. *Source: Author's dataset*

Table 10: Estimation of Value Added Per Worker at the Firm Level: All Firms

	(1)	(2)	(3)	(4)
	Ln(Population)	**Ln(Landscan)**	**Ln(Employment 2002)**	**Ln(Own Employment 2002)**
Ln(Own-Cell)	0.0934	0.0883	-0.0778***	-0.0637***
	(0.0675)	(0.0541)	(0.0242)	(0.0190)
Ln(Ring 1)	-0.237**	-0.0863	-0.0765	0.0140
	(0.118)	(0.112)	(0.0493)	(0.0335)
Ln(Ring 2)	-0.00309	-0.0190	0.174**	0.125***
	(0.129)	(0.127)	(0.0734)	(0.0448)
Ln(Ring3)	0.179*	-0.0129	-0.0199	-0.0381
	(0.100)	(0.0903)	(0.0632)	(0.0444)
Constant	8.101***	8.807***	8.255***	8.450***
	(0.755)	(0.799)	(0.801)	(0.694)
Observations	2342	2342	2342	2342
R-squared	0.082	0.079	0.087	0.086

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of value added per worker for each firm observation. Controls are dummies for the industry sector each firm is in, distance to the city center (km), distance to lake Victoria (km), and distance to the Kampala Northern Bypass Road (km). Robust standard errors are given in parentheses. Asterisks indicate $p < 0.01(***)$, $p < 0.05(**)$, and $p < 0.1(*)$. *Source: UBI Firm Survey.*

Table 11: Estimation of Value Added Per Worker at the Firm Level: Manufacturing and Business Services Firms

	(1)	(2)	(3)	(4)		(5)	(6)
	Ln(Population)	**Ln(Landscan)**	**Ln(Employment 2002)**	**Ln(Own Employment 2002)**		**Ln(Own-Employment 2002)**	**Ln(Own Firms 2002)**
Ln(Own-Cell)	0.0823	0.156	-0.158***	-0.0287	0-1km	-0.0978**	-0.163***
	(0.143)	(0.128)	(0.0513)	(0.0374)		(0.0442)	(0.0454)
Ln(Ring 1)	-0.579**	-0.356	-0.241**	0.0901	1-2km	-0.0775	-0.0284
	(0.256)	(0.285)	(0.117)	(0.0847)		(0.0605)	(0.0662)
Ln(Ring 2)	0.433	0.447	0.412**	0.407***	2-3km	0.263***	0.186**
	(0.278)	(0.315)	(0.175)	(0.109)		(0.0764)	(0.0783)
Ln(Ring3)	0.113	-0.317	0.0326	0.260**	3-4km	-0.0734	0.0125
	(0.230)	(0.222)	(0.162)	(0.121)		(0.0688)	(0.0757)
Constant	9.092***	10.21***	8.780***	4.368***	Constant	5.185***	5.073***
	(1.157)	(1.229)	(1.301)	(0.861)		(0.348)	(0.291)
Observations	534	534	534	534	Observations	525	525
R-squared	0.079	0.066	0.107	0.143	R-squared	0.046	0.059

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of value added per worker for each firm observation. Controls are dummies for the industry sector each firm is in, distance to the city center (km), distance to lake Victoria (km), and distance to the Kampala Northern Bypass Road (km). In columns 1-4, we provide results where the independent variables are calculated from rings made up of data in 1km grid cells; whereas, in columns 5-6, results are based on data within a circular neighborhood of each firm. Robust standard errors are given in parentheses. Asterisks indicate $p < 0.01(***)$, $p < 0.05(**)$, and $p < 0.1(*)$. *Source: UBI Firm Survey.*

Figure 1: Kampala Densities

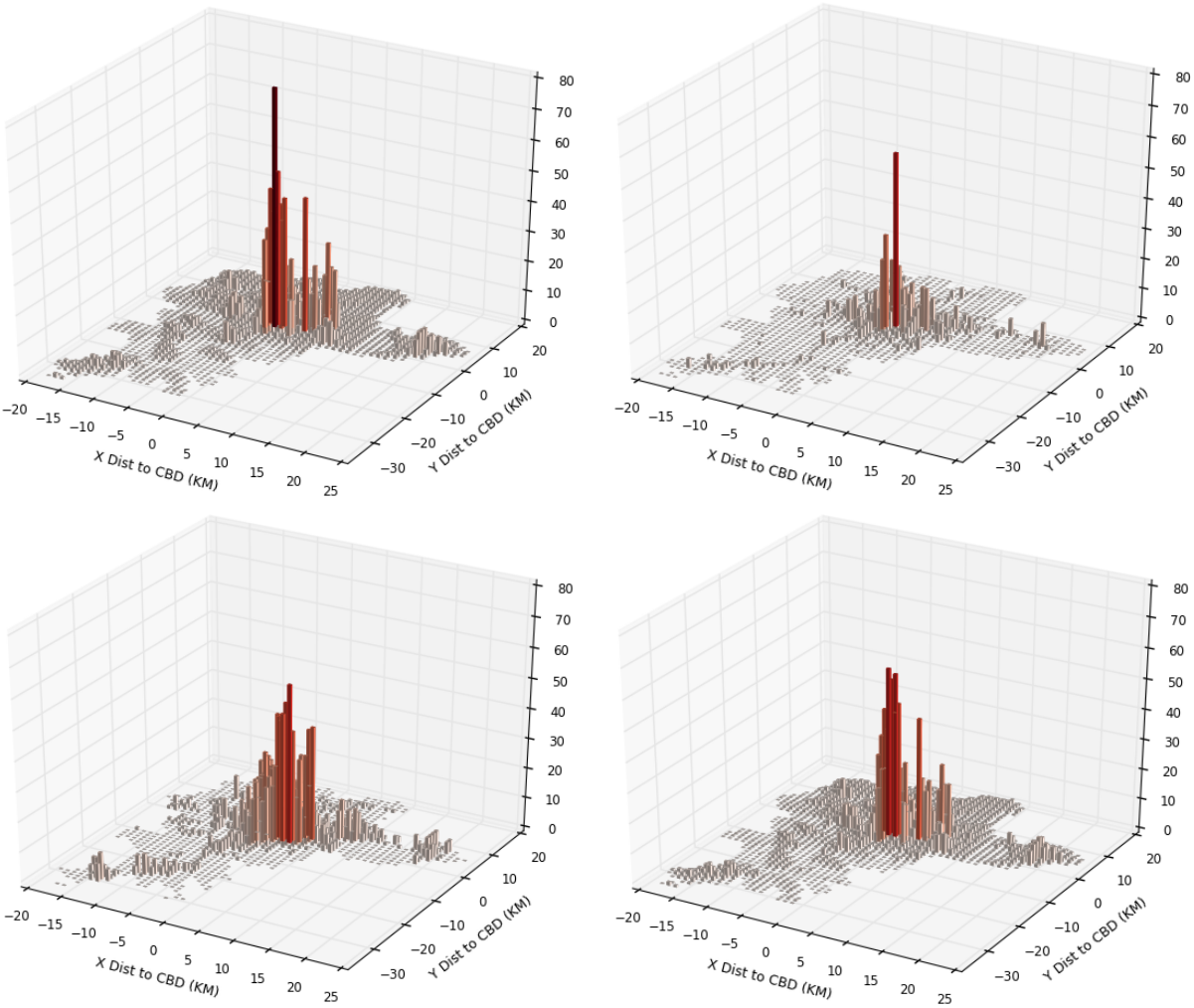


Figure 2: Nairobi Densities

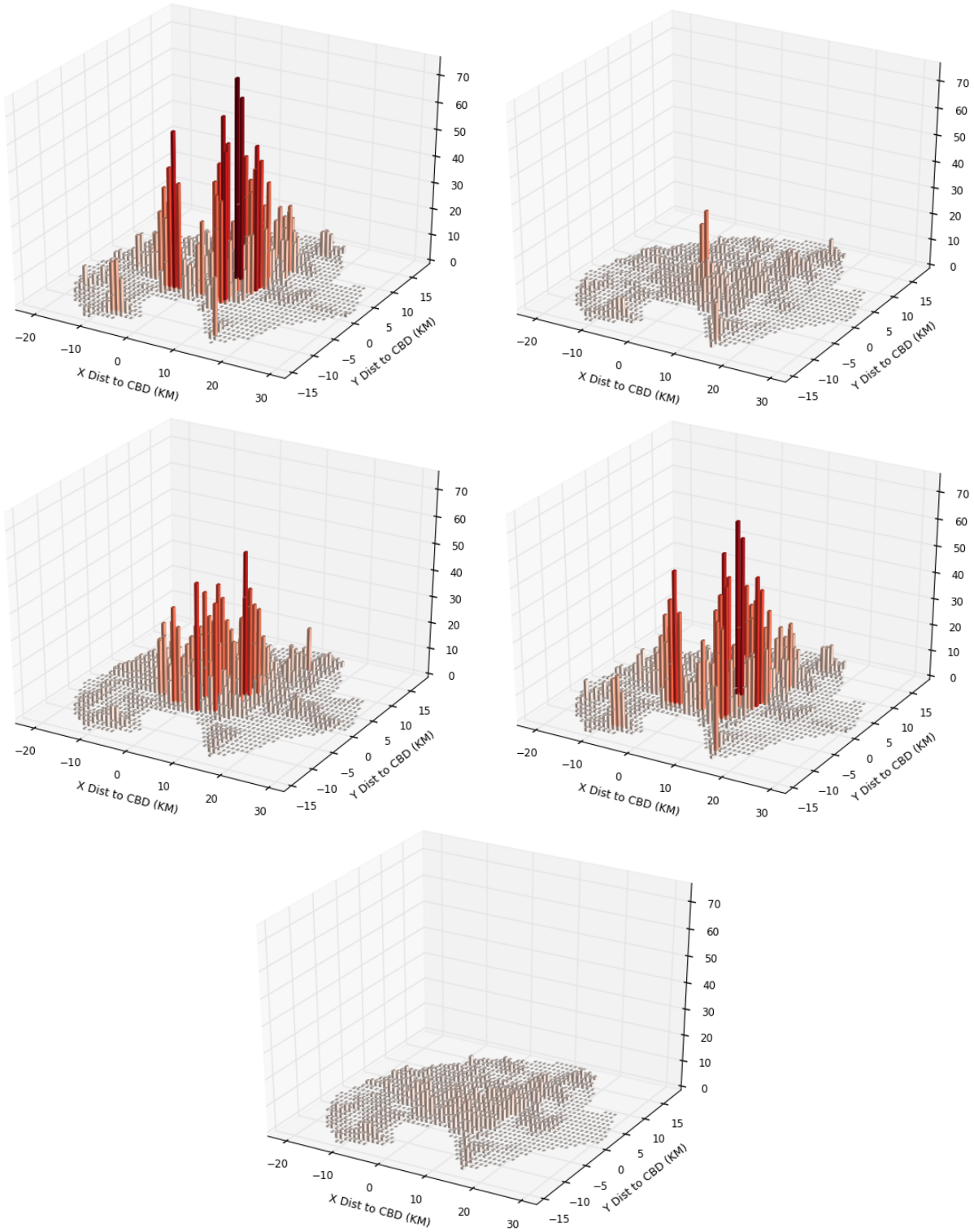


Figure 3: Defining Cities and Settlements

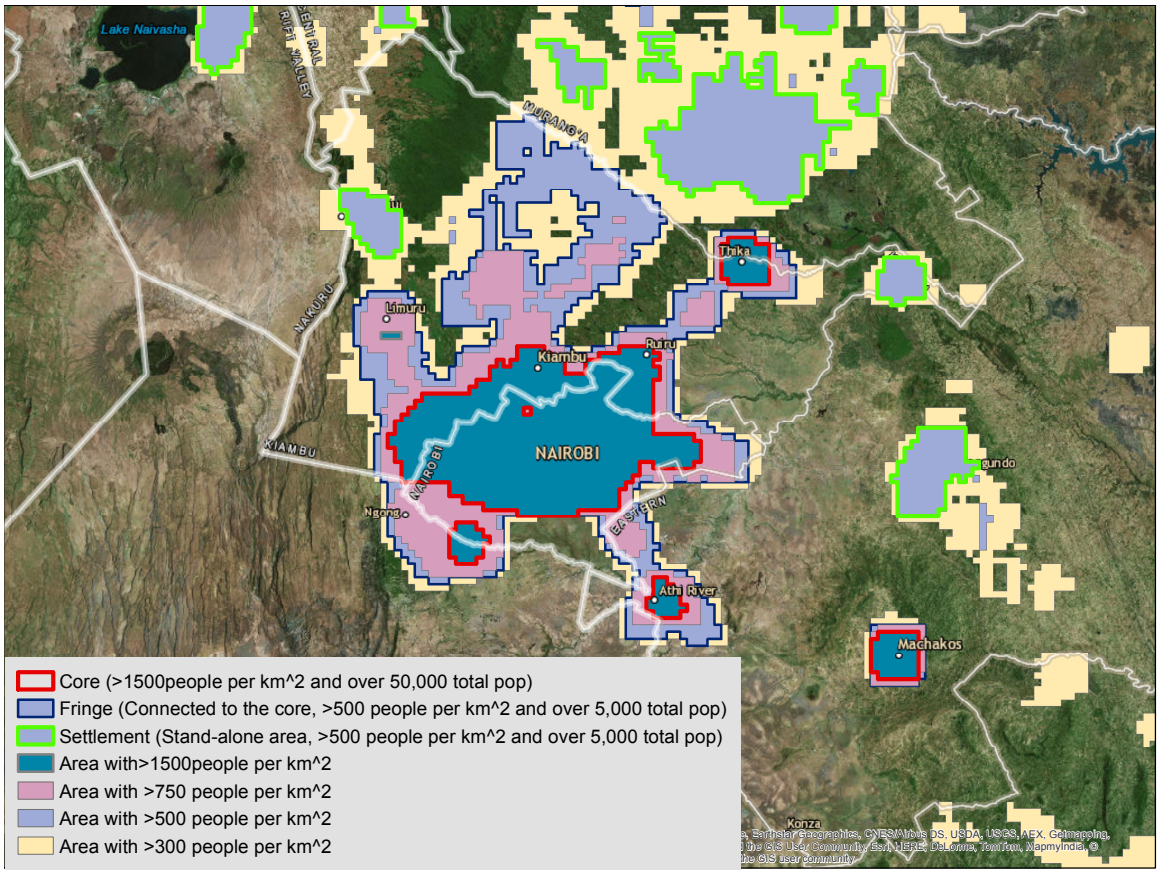


Figure 4: Differences in City Layout and Density Measures

City 1

5	5	5	5	5	5
5	5	5	5	5	5
5	5	5	5	5	5
5	5	5	5	5	5
5	5	5	5	5	5
5	5	5	5	5	5

City 2

0	0	0	10	10	10
0	0	0	10	10	10
0	0	0	10	10	10
0	0	0	10	10	10
0	0	0	10	10	10
0	0	0	10	10	10

City 3

0	10	0	10	0	10
10	0	10	0	10	0
0	10	0	10	0	10
10	0	10	0	10	0
0	10	0	10	0	10
10	0	10	0	10	0

Figure 5: Average City PPD by Countries Around the World

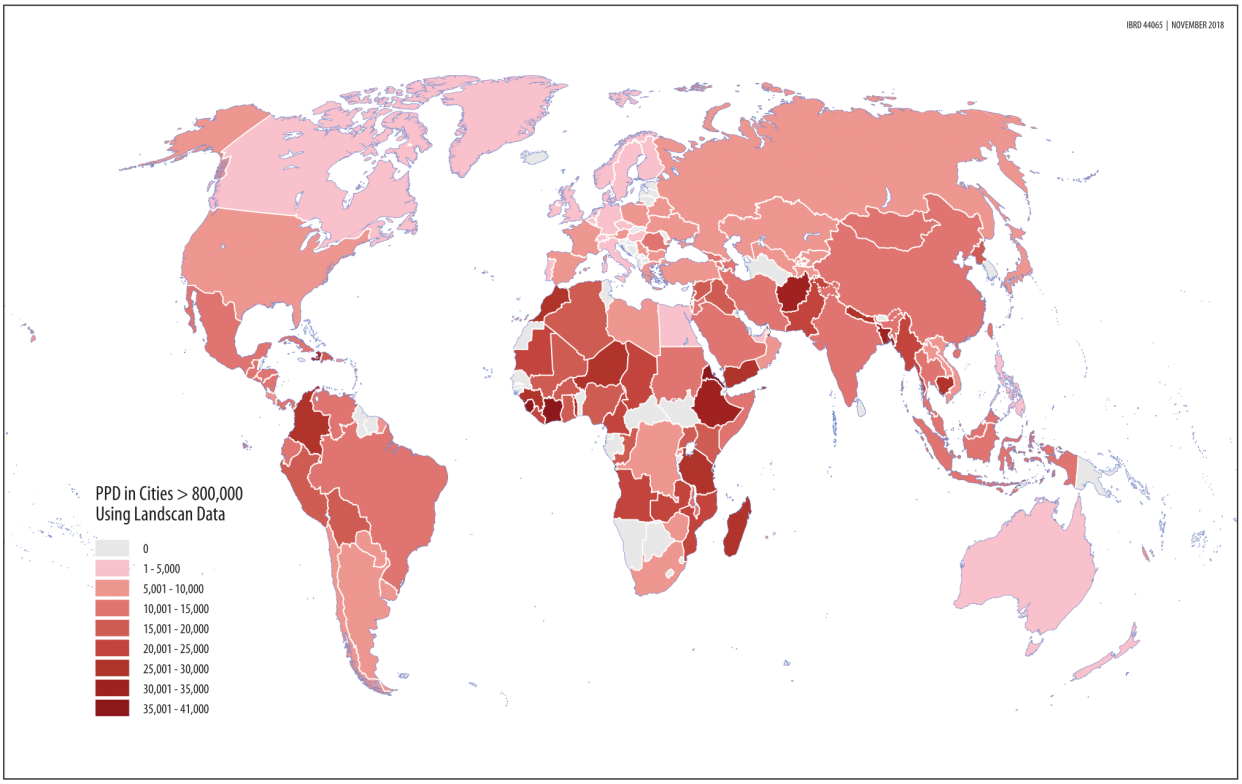


Figure 6: Overall Size Distribution of Urbanized Areas

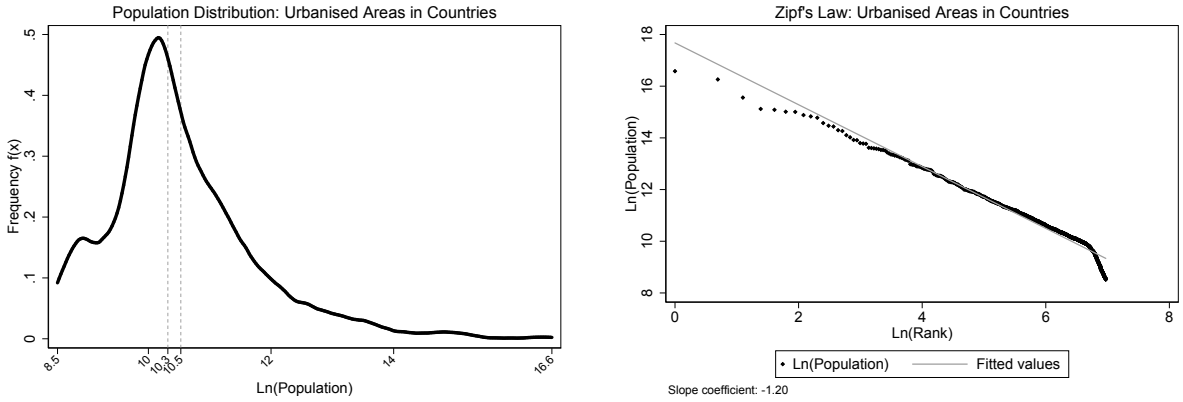


Figure 7: Size Distribution of Urbanized Areas

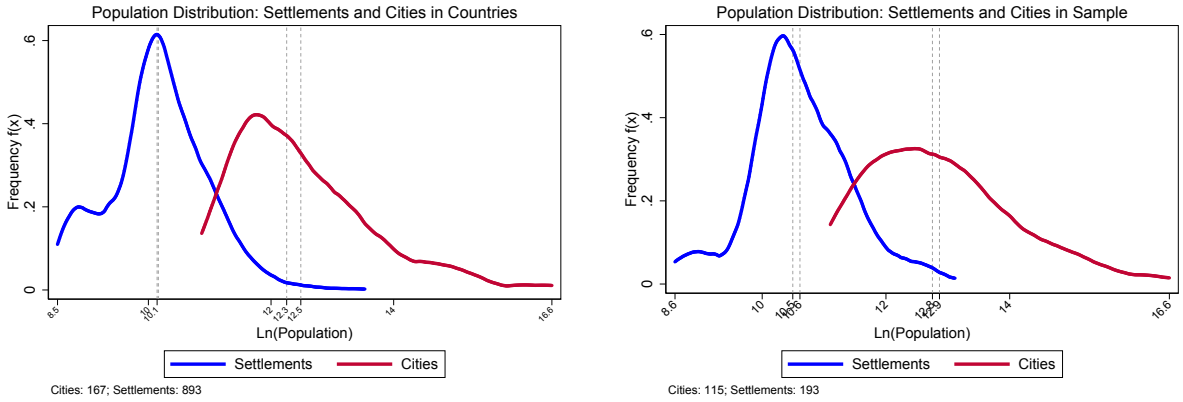


Figure 8: Optimisation of RPA Discount Rate

Appendix A: Results

Table A1: All Results

	Ln(Population)			Ln(PD)			Ln(PPD)			Ln(RPA)		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Rural X Scale	0.00150	-0.0148	-0.0148	-0.00833	-0.0213*	-0.0198*	0.0309***	0.0157*	0.00193	0.00276	-0.0123	-0.0136
	(0.0112)	(0.0108)	(0.0104)	(0.0115)	(0.0112)	(0.0107)	(0.00850)	(0.00825)	(0.00783)	(0.0111)	(0.0107)	(0.0103)
Urban X Scale	-0.0861**	-0.0827**	-0.0845***	0.782***	0.524***	0.426***	0.365***	0.195***	0.107**	0.596***	0.302***	0.196**
	(0.0351)	(0.0327)	(0.0310)	(0.167)	(0.155)	(0.147)	(0.0582)	(0.0546)	(0.0521)	(0.0964)	(0.0909)	(0.0865)
City X Scale	0.100***	0.144***	0.167***	-0.0588	0.0305	0.108	0.213***	0.254***	0.327***	-0.107	0.0731	0.164*
	(0.0372)	(0.0346)	(0.0329)	(0.169)	(0.157)	(0.148)	(0.0627)	(0.0584)	(0.0558)	(0.0982)	(0.0923)	(0.0879)
Urban	1.471***	1.090***	0.984***	-4.658***	-3.198***	-2.669***	-2.242***	-1.134**	-0.625	-5.190***	-2.653***	-1.764**
	(0.394)	(0.368)	(0.349)	(1.103)	(1.025)	(0.966)	(0.476)	(0.445)	(0.424)	(0.935)	(0.881)	(0.838)
City	-0.583	-1.376***	-1.763***	0.276	-0.306	-0.871	-1.874***	-2.198***	-2.849***	0.773	-0.921	-1.842**
	(0.417)	(0.388)	(0.368)	(1.118)	(1.038)	(0.979)	(0.520)	(0.484)	(0.462)	(0.955)	(0.897)	(0.853)
Observations	44157	44140	44140	44157	44140	44140	44135	44118	44118	44157	44140	44140
R-squared	0.254	0.332	0.386	0.265	0.338	0.391	0.264	0.337	0.390	0.265	0.337	0.390
Country-Year FE	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Controls		✓	✓		✓	✓		✓	✓		✓	✓
Occupation FE			✓			✓			✓			✓

Notes: Table reports results from OLS regressions; the dependent variable is the natural logarithm of total net household income from all available sources. Controls are the education (recorded or not), level of education (if recorded), age, age squared and gender of the household head; household size; household size squared; whether the household owns land and, if so, the natural logarithm of the size of land holdings in hectares. Each column includes country-year fixed effects. Standard errors are clustered at the individual urbanized area level. Asterisks indicate $p < 0.01(***)$, $p < 0.05(**)$, and $p < 0.1(*)$. *Source: World Bank Living Standard Measurement Surveys.*

Table A2: Summary Statistics of Full Sample

Variable	Mean	Min	Median	Max	STD	N
Rural X Local Scale Measures						
Ln(Own-Cell Population)	2.83	0.00	3.40	8.96	2.42	44140
Ln(PD)	3.09	0.00	4.16	7.12	2.34	44140
Ln(PPD)	3.99	0.00	5.24	9.89	3.04	44118
Ln(RPA)	4.74	0.00	6.63	9.42	3.49	44140
City X Local Scale Measures						
Ln(Own-Cell Population)	1.95	0.00	0.00	11.14	3.44	44140
Ln(PD)	2.02	0.00	0.00	10.17	3.45	44140
Ln(PPD)	2.38	0.00	0.00	10.92	4.05	44140
Ln(RPA)	2.70	0.00	0.00	12.79	4.59	44140
City X City-Wide Scale Measures						
Ln(Total Population)	3.66	0.00	0.00	16.58	6.21	44140
Ln(PD)	1.94	0.00	0.00	8.65	3.29	44140
Ln(PPD)	2.42	0.00	0.00	10.81	4.09	44140
Ln(1 + CV Term)	0.47	0.00	0.00	3.04	0.82	44140
Ln(RPA)	2.83	0.00	0.00	12.28	4.79	44140
Ln(AD)	2.57	0.00	0.00	11.13	4.34	44140
Ln(1 + Cov Term)	0.26	0.00	0.00	1.88	0.46	44140

Table A3: Summary Statistics of City Sample

Variable	Mean	Min	Median	Max	STD	N
City X Local Scale Measures						
Ln(Own-Cell Population)	7.51	0.00	7.68	11.14	1.94	11493
Ln(PD)	7.77	5.77	7.64	10.17	1.07	11493
Ln(PPD)	9.16	6.34	9.29	10.92	0.97	11493
Ln(RPA)	10.38	8.22	10.28	12.79	1.10	11493
City X City-Wide Scale Measures						
Ln(Total Population)	14.05	11.10	14.44	16.58	1.43	11493
Ln(PD)	7.47	6.51	7.34	8.65	0.61	11493
Ln(PPD)	9.29	7.36	9.29	10.81	0.71	11493
Ln(1 + CV Term)	1.82	0.81	1.76	3.04	0.37	11493
Ln(RPA)	10.87	9.26	10.62	12.28	0.84	11493
Ln(AD)	9.87	8.95	9.70	11.13	0.60	11493
Ln(1 + Cov Term)	1.00	0.26	0.99	1.88	0.31	11493

Appendix B: Data Description and Methodology

Overview

The following sections provide further details on the data sources and methodologies for data preparation and usage in this paper. Section B.1 and its sub-parts describe the LandScan (2012) data and how we use it to construct urbanized area boundaries, as well as why we chose certain density cut-off criteria for the construction of the boundaries. Section B.2 describes the Living Standards Measurement Surveys and finally Section B.3 describes how we harmonize the two datasets to study how different density measures relate to wage premia.

B.1 Landscan

For our analyses in Sections 3-6 we make use of a global dataset of population density developed by Oak Ridge National Laboratory. Their dataset, LandScan (2012), estimates population density at approximately 1km resolution (30″ X 30″ arc-seconds) for the entire globe. The data represent an ambient population (average over 24 hours) and have been developed to manage global disaster relief. Population density is estimated by combining satellite imagery detailing the extent of built area within a certain location, with population census data at an administrative level such as the county or parish level. The final population estimates for each grid square are determined by an algorithm that allocates population, with some weights, from the administrative level to the finer grid level based on built cover. This methodology reflects the fact that people are likely to occupy other buildings in the city over the course of a day for work and recreational purposes, as opposed to consistently staying in the residential areas where they answered the census questionnaire.

A major limitation of the Landscan data is that although the built cover information is consistent, since it is derived from satellites, the census sub-units used may vary across countries and may be large. At times Landscan uses other data to improve accuracy, such as topographic maps or roads, but that also depends on the data available for each city. Finally, the whole procedure of how the data is created and what weights are used to

assign population is somewhat of a black box. Nevertheless, this is the best source of globally consistent cross-sectional population data are available to date.

B.1.1 Using Landscan to Create Urbanized Area Boundaries

We use Landscan to define three unique types of urban areas: cores, fringes, and low-density settlements. A city is the consolidation of its core and surrounding fringe, whereas a settlement is a unique and separate urbanized area. All of these urbanized area boundaries are defined based off density thresholds that we assign to each unique 1km grid-square in Landscan, followed by total population thresholds that we assign to the contiguous sets of cells to determine whether they qualify as an urban core, fringe, or LD settlement. As we are only using Landscan, these boundaries make no use of administrative borders to define urban extents.

To calculate the density of each grid-cell, we apply a smoothing methodology where each reference cell is assigned the average density from a neighborhood which includes itself and a 7x7km square around it, where the reference cell falls in the center of the 7x7km neighborhood square. This approach is essential for dealing with natural breaks in density in the data which may relate to changes in land use, terrain, and building restrictions within urban areas. Consider, for instance, Central Park in Manhattan, or the River Thames in London; assigning a density criterion just to each singular grid-cell would lead to unnatural breaks/holes in our urban area polygons which would challenge our ability to analyze our urban areas as singular units.

Once we have calculated the average density of each grid-cell within its 7x7km neighborhood, we consolidate all contiguous grid cells that have an average density above certain thresholds, where the thresholds vary depending on the type of urban unit we are creating: urban core, fringe, or settlement. Contiguous cells are combined based on a rook neighbor relationship, wherein, the rook neighbors are the four cells adjacent to the reference cell in the vertical or horizontal direction (as outlined in Figure B1a), but not including the diagonally adjacent cells which are queen neighbors (as outlined in Figure B1b). Finally, contiguous cells meeting certain density criteria will be classified as a city (including the urban core and fringe) or an isolated LD settlement, if all the cells meet the average density

Figure B1a: Rook Neighbor

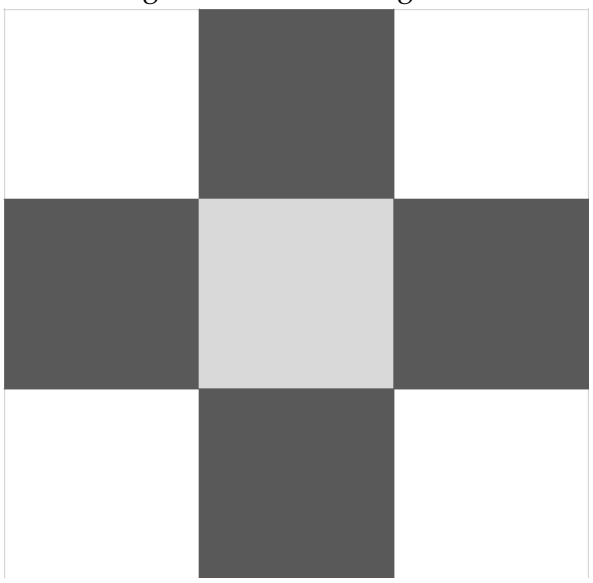
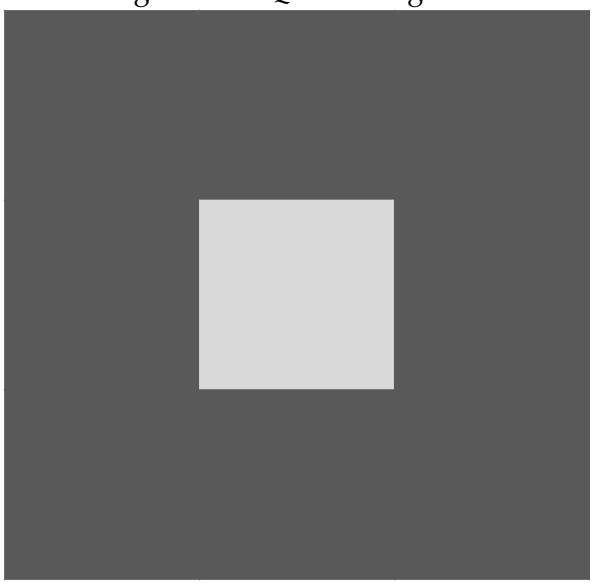


Figure B1b: Queen Neighbor



cutoffs, and the population added up over the contiguous cells is above a certain threshold.

Specific information on the thresholds we apply for density and total populations for our urban areas is detailed in the sections below. First, for the initial exercise in Section 3 where we compare African cities above 800,000 to other cities in the world above 800,000. Then for the African cities and settlements polygons which are used in the remainder of the paper to study city size characteristics and wage premiums.

B.1.2 Calculating Metropolitan Areas for the World

In Section 3, we create polygons for cities with a population above 800,000. These metropolitan areas are defined by the following criteria: first, each grid cell in the metropolitan area must have an average density of 1,500 people per sq. km. or more; second, each polygon

must contain at least one metropolitan area of above 300,000 people as defined by the UN (2015); third, the sum of the UN metropolitan area or areas within our polygon must be at least 800,000 people.

We use the intersection with the UN metropolitan areas data as it provides an external check on the Landscan data. One issue we noticed in India and China, in particular, was that large swathes of seemingly high density rural areas had been combined into gigantic urban areas, which we were certain were not urban areas in reality. Presumably this was occurring because Landscan is smearing people in these regions into agricultural areas. Hence, the UN allows us to sense check the data against reported evidence.

Applying our initial density-cutoffs to create contiguous polygons, leaves us with a set of 13,638 metropolitan area polygons. After we overlay this with the coordinates of the metropolitan areas above 300,000 in the 2015 UN data, we get 1,544 polygons which match and we therefore keep. Only 54 of the total 1,692 UN points above 300,000 did not match with our polygons. In the case that several metropolitan areas fall into one polygon, the UN population is summed and the name of the largest metropolitan area is kept. Once we apply the total population threshold of above 800,000, we are finally left with a set of 599 cities worldwide, with 451 in the developing world.

B.1.3 Creating Core, Fringe and Settlement Polygons for Africa

For separate analyses which we do just on Africa in Sections 4-6, we define urbanized areas using similar but distinct density and population threshold cutoffs. We define three unique urbanized areas: cores, fringes, and LD settlements. A city is the consolidation of an urban core and its surrounding peripheral fringe, whereas a settlement is a unique area. We define each area as follows:

1. Cores are defined as:

    - Contiguous pixels of density of above 1,500 people per sq. km. (rook neighbors), where the set of contiguous cells has a total summed population of at least 50,000 according to Landscan

2. Fringes are defined as:

- An attached set of peripheral and contiguous adjoining cells around the urban core which each have an average density above 500 people per sq. km.

3. Cities are defined as:

   - The consolidation of the core and fringe areas defined above.

4. Settlements are defined as:

   - Contiguous pixels of density of above 500 people per sq. km., where the set of contiguous cells has a total summed population of at least 5,000 people according to Landscan. These areas are 'stand-alone' and therefore, separate to cities as defined above.

B.2 Living Standards Measurement Surveys (LSMS)

The Living Standards Measurement Surveys have been conducted in a number of developing countries by the World Bank and the national statistical offices of the country in question. To study wage premia in this paper, we make use of surveys for Ethiopia, Ghana, Malawi, Nigeria, Tanzania, and Uganda. All of these surveys are considered representative of households at the national level, as well as urban/rural and major ecological zones of the countries. In Table B1 below, we provide a list of the surveys we use and the number of households surveyed in each of these rounds.

In each sample, a two-stage probability sampling methodology is used. In the first stage, Primary Sample Units (PSUs) are selected based on the probability proportional to size of all of the enumeration areas in geographic zones in the country. In the second stage, households are then selected randomly from each PSU, after which, each individual within a household is surveyed. All of the LSMS surveys are publicly available for download from the World Bank website, so for further information on any individual survey and its methodology, we refer the reader to the information documents provided by the World Bank.

Although the contents of each survey vary, they all have quite consistent data at the household and individual level on aspects such as income, educational attainment, demographics, labor allocation, asset ownership and dwelling characteristics, as well as geo-

Table B1: LSMS Surveys

Country	Survey	Year	Sample Size
Ethiopia	Socioeconomic Survey	2011	3,917
		2013	5,073
		2015	5,263
Ghana	Socioeconomic Panel Survey	2010	4,662
		2013	4,634
Malawi	Integrated Household Survey	2010	3,246
		2013	3,104
Tanzania	Panel Household Survey	2008	3,280
		2010	3,924
Uganda	National Panel Survey	2009	3,123
		2010	2,716
		2011	2,716
		2012	3,119
Nigeria	National Household Survey	2010	5,000
		2012	5,000

graphical identifiers locating the latitude and longitude of the centroid of each enumeration area.

Agricultural households report on various aspects of farming such as crop choice, inputs on the farm, labor usage and the types of land allocation such as harvesting, grazing or fallow. Among non-agricultural households, additional modules are provided on whether they are self-employed with their own business and if so the revenues and various factor costs of that business. In some cases, aggregation of revenues and costs at the household level is already computed in the survey and these aggregations are used where possible. For example, labor income at the individual level is already aggregated in the surveys to include all wage, in-kind and bonus income from all jobs. Elsewhere, input costs of agricultural and non-agricultural businesses are aggregated to the household level.

We calculate income from the survey data and aggregate either to the individual or household level (depending on our analysis) using all available sources of money flowing in. Letting i index an individual or household, this can be summarised as follows:

$$Y_i = \Sigma_i Y_i^{SE} + \Sigma_i^L + \Sigma_i^K \tag{5}$$

where y_i^{SE}, y_i^L, and y_i^K represent self-employed income, labor income, and capital income respectively. Households reported receipts of incomes through various forms and over various time intervals. The variables used for income receipts and the time intervals over which they were received are reported as follows:

Table B2: Income sources and time intervals in LSMS surveys

Income Source	**Time Interval**
Last payment in cash	Hour, Day, Week, Fortnight, Month, Quarter or Year
Last payment in kind (value in LCU)	Hour, Day, Week, Fortnight, Month, Quarter or Year
Net income from business	Week or Month
Remittances in cash	Year
Remittances in kind	Year
Rent of property	Year
Private or govt pensions	Year
Domestic remittances	Year
Rent of farmland	Year or cropping season
Sales of crops	Year or cropping season
Sales of crop residue	Year or cropping season
Sale of livestock products	Year or cropping season

All revenues are converted to a monthly interval. In cases where incomes are reported over the year, quarter, fortnight, or week, the variables are scaled to a monthly value simply by multiplying by the ratio of a month to the time interval in question (for instance a quarter is multiplied by 1/3 to be monthly). In cases where the last income payment is reported based on a day of work, the figure is multiplied by the average days the respondent reports to work each week, and then multiplied by 52/12 to be a roughly monthly figure. In cases where the last income payment is reported based on an hour of work, the figure is multiplied by the average hours the respondent reports to work each day, then the average days they report to work each week, and finally by 52/12 to get a roughly monthly figure.

A similar method is used to convert expenses to a monthly aggregate figure. The reported expenses in the household surveys are as follows:

We conduct analyses both at the household and individual level, subtracting expenses from revenues to get our net income figure. For the individual analysis we choose to restrict our sample to the set of individuals in households that do not own any agricultural land because farm income was recorded at the household level in the surveys and thus, we

Table B3: Expenses and time intervals in LSMS surveys

Expense	**Time Interval**
Wages	Month
Raw materials	Month
Other expenses	Month
Farm inputs	Year
Additional agricultural expenses	Year

have no way to assign farm income to individuals. We look at adults aged 18-65 who are working part or full time with income in our analysis at the individual level.

At the household level, we construct measures of income including income from self-employment, labor, capital, and land income. All income is measured before taxes. We choose to include in the calculation of monthly household income, transfer payments such as remittances, gifts and pensions and we subtract transfer payments flowing out of the households. These sources of incomes are likely to be important for the budget constraints of households, particularly in rural communities, so they are important in our study of urban-rural income disparities, although we note that our results are robust to excluding remittances.

B.3 Harmonizing the Data with Our Urbanized Areas

Fortunately, the LSMS surveys provide latitude and longitude coordinates of enumeration areas in the sample for each country. This allows us to directly harmonize the surveys with our data from Landscan based on their spatial relationship. For each survey area that is assigned to an urbanized area from our created polygons, they are assigned urban area wide density measures as well as local density measures which we calculate based off the neighborhood within approximately a 6km radius around the grid-cell the household is assigned to. For rural areas, which are not assigned to any urbanized areas, we only assign them to an own-cell density measure from the cell they are located within, as well as a local neighborhood density measure, again based off the neighborhood of radius approximately 6km around the cell the household is assigned to.

Figure B4a: Density Ring Measures