CSAE Working Paper WPS/2022/11

Informal Settlements & Consumption Gaps: Decomposing the Urban-Rural Consumption Gap Within African Countries

Emma Buckland*

September 17, 2022

Abstract

Regional inequality, epitomised by the urban-rural consumption gap, is considerable in the developing world. I use a city-based approach to decompose the gap in ten sub-Saharan African countries, evaluating living standards by proximity to cities (in rural areas) and size of cities (in urban areas). I further decompose the consumption gap by incorporating urban informal settlements ('slums'). Despite the prominence of slums in sub-Saharan Africa, they are under-studied as they are difficult to identify and connect with survey data. I address those challenges by (i) creating the first transcontinental map of slums; and (ii) improving upon current best practice in connecting spatial and survey data in sub-Saharan Africa. Within slums, I proxy for the formality of housing by creating a tool that measures how regularly – on orthogonal axes – buildings are laid out. To measure living standards, I implement the approach of the seminal paper measuring consumption gaps, Young (2013), via Item Response Theory. I use the detailed regional decomposition of living standards to reevaluate potential explanations for the urban-rural divide.

*I thank Simon Quinn and Douglas Gollin for their comments and insight.

Centre for the Study of African Economies
Department of Economics · University of Oxford · Manor Road Building · Oxford OX1 3UQ
T: +44 (0)1865 271084 · F: +44 (0)1865 281447 · E: csae.enquiries@economics.ox.ac.uk · W: www.csae.ox.ac.uk

1 Introduction

The urban-rural consumption gap comprises a substantial share of inequality in the developing world. The seminal paper measuring the gap across countries, Young (2013), estimates that it accounts for approximately 40% of total inequality[1] in the nearly 70 developing countries included in the study (Young, 2013; Lagakos, 2020). The urban-rural gap is also associated with (and similar in magnitude to) the agricultural productivity gap, with the lack of movement from rural, agricultural areas to urban ones being a potential obstacle to structural transformation (Gollin et al., 2014). Consequently, understanding the urban-rural living standards gap – where I proxy living standards by consumption throughout this paper – may shed light on both growth and inequality in the developing context.

However, the inability to measure living standards in urban informal settlements – otherwise known as 'slums' – has hindered further exploration of the urban-rural divide. With rapid urbanisation, slums in sub-Saharan Africa have been expanding in population by 4.5% per annum. This rate implies the resident population doubles every fifteen years. Rural-urban migration comprises a substantial portion of this expansion (Marx et al., 2013; UN-Habitat, 2013). Marx, Stoker, and Suri (2013) argue, based on a rare cross-country study of slums in four developing countries, that millions remain in persistent poverty in informal settlements. Accordingly, they contend that slums form a kind of 'poverty trap' that prevents residents from accessing the benefits of the city.

Despite the prominence of slums in sub-Saharan Africa, conditions in informal settlements have been under-studied. This may be due to the following challenges: (i) there are limited maps of slums, and no publicly available cross-country ones to my knowledge; and (ii) linking large-scale anonymised survey data with slum locations has been considered infeasible (Gollin et al., 2021), even if those slum locations were known. This paper addresses both these obstacles to evaluate living standards in slums across ten sub-Saharan African countries comprising approximately half the total population of sub-Saharan Africa.

[1] According to the measure of inequality developed in Young (2013), which will be discussed in detail in later sections.

To address the first challenge, I create the first transcontinental map of informal settlements. Informal settlements are notoriously difficult to map, mainly due to accessibility constraints. I overcome this by using a recently released map of building footprints (made available by Google Research in 2021). Using a dataset based on satellite imagery rather than surveys largely overcomes the accessibility constraint. The resolution of the imagery is sufficient to identify the small buildings that characterise informal settlements. For some countries, the map created in this paper seems to be the first map of informal settlements at all.

In the process of identifying slums by their 'geographic footprint' of small, dense buildings (Michaels et al., 2021), I process the building footprints data to create two types of additional datasets: mean base/footprint area and building density maps, both spanning the entirety of the land mass of sub-Saharan Africa (except Cameroon and South Sudan due to data limitations). I repeat much of the analysis for different subsets of the building footprints, in order to find the measures for identifying slums that are most robust to the modeler's choices – and so facilitate geographical consistency.

Aside from identifying the location of slums at scale for the first time, I also attempt to measure inequality within slums. I create a tool that quantifies how regularly laid out buildings are (as a proxy for planning), since those living in the most informal housing may have particularly low living standards. This tool calculates the standard deviation of the angles of building walls (in an area) from two orthogonal axes, where the axes are chosen to minimise that deviation.[2] I demonstrate I can obtain this measurement at scale (for over half a million $100m^2$ areas in cities). The tool I develop is also applicable to other spatial questions in economics. For example, how the arrangement of a city's roads relates to its economic prosperity, or in general as a proxy for city planning.

Once slums are identified, the second challenge to measuring living standards is connecting slum areas to spatial data. The Demographic & Health Survey (DHS) dataset – one of the only nationally representative cross-country surveys in sub-Saharan Africa – displaces respondent locations in urban areas (up to 2kms)

[2] The axes are the circular mean and the line perpendicular to it, as will be described later.

to preserve anonymity. This makes finer distinctions between neighbourhoods in urban areas challenging to make (Gollin et al., 2021). However, this paper improves upon current best practice in connecting DHS observations to spatial characteristics. I develop a tool that uses probability weighting to achieve this. Ultimately, I am able to identify households across sub-Saharan Africa likely to live in informal settlements, a task previously considered infeasible (Gollin et al., 2021). This more sophisticated probability weighting has uses beyond this paper. For example, this technique could be used to map access to amenities within a city, to proxy living standards and perhaps to effectively allocate resources.

Many of the tasks described so far are made computationally feasible at scale due to the extensive use of parallelisation (with cloud computing). To implement parallelisation, I use the Google Earth Engine Platform (distinct from Google Earth). The use of Earth Engine is rare in spatial economics, but increasingly common in other spatial fields. While the platform has fewer built-in functions useful to economists compared to conventional platforms like QGIS or ArcGIS, it has processing advantages as it automatically uses the Google Cloud Platform (cloud computing) to implement tasks simultaneously. These processing power gains would be non-trivial to implement in ArcGIS/QGIS. Among other things, the use of parallelisation allows information-value rather than computational limitations to take priority in the method, facilitating the development of multiple datasets (described previously) with uses beyond this paper.

Once slums are identified and connected to survey responses, I create a consumption index to measure living standards. The seminal paper measuring consumption gaps in developing countries is Young (2013). I bring to light the similarities between Young's approach and a simpler model: Item Response Theory (IRT). I develop a methodology that implements Young's technique as a modification of an IRT model. This implementation increases the transparency and accessibility of the method - which has uses beyond this paper as it may allow this adapted Young approach to compete with other commonly used consumption indices like principal component analysis.

The aforementioned innovations make it possible to explore regional inequality in

finer detail than was previously possible. For example, Young (2013) only measures average living standards for urban and rural areas. This paper decomposes the urban-rural consumption gap in ten sub-Saharan countries, and exposes welfare distribution within rural and urban areas. This distribution is core to understanding the cause of the consumption gap. Before discussing the specific findings of this paper, it is worth considering the current prominent explanations for the urban-rural divide: frictions, sorting and spatial equilibrium.

Frictions may prevent migration to cities, creating inequality between regions (a 'geographic poverty trap') and inefficient labour allocation (Kraay and McKenzie, 2014; Lagakos, 2020).[3] Additionally/alternatively, the urban-rural disparity could be due to sorting if higher-ability individuals disproportionately move to urban areas.[4] Empirically, large-scale panel studies in Kenya, Indonesia and Brazil have found that migrants experience substantially lower wage gains than the cross-sectional consumption gaps would suggest (Hicks et al., 2021; Alvarez, 2020). This has been used as evidence in favour of sorting over frictions as an explanation for the urban-rural gap (Lagakos, 2020).

A further potential explanation is the existence of a spatial equilibrium with the urban-rural gap reflecting differing preferences for rural compared to urban life. However, Gollin et al. (2021) illustrate that indicators of quality of life[5] tend to increase monotonically with population density, by quartile, in 21 sub-Saharan countries. By contrast, a spatial equilibrium would imply trade-offs/substitution between indicators for utility across households to be equalised.

Finding the welfare distribution within urban and rural areas, as this paper does, contributes to understanding the causes of the gap in two main ways: Firstly, if there is inequality within areas, moving for some could be less

[3] In the frictions paradigm, the movement of workers out of agriculture and into other sectors (more prominent in cities) has the potential to increase welfare and productivity (McMillan et al., 2014; Restuccia et al., 2008; Vollrath, 2014; Caselli, 2005).

[4] For example, there is ample evidence that sorting occurs (away from agriculture) based on education. Cross-country analysis of over 120 countries showed that agricultural workers, on average, have substantially less education than their counterparts (Gollin et al., 2014). Similarly, when observing consumption gaps (net of education), unobservables like ability could be driving the discrepancy.

[5] The measures of quality of life included public goods, asset ownership, crime and pollution.

transformational than moving for others, even in the frictions (rather than the sorting) paradigm. In other words, it would matter where in a rural area to where in an urban area a migrant moves. Secondly, informal settlements may be a bottleneck for migration if living standards are sufficiently low. Perhaps for this reason, Gollin et al. (2021) acknowledge that the monotonic relationship between population density and quality of life indicators may not hold for informal settlements. If monotonicity breaks down, a spatial equilibrium (between rural areas and urban slums) once again becomes a viable explanation for the urban-rural gap – although there still may not be spatial equilibrium within cities.

To explore these avenues, I first use a city-based approach to decompose the consumption gap in ten sub-Saharan African countries, evaluating living standards by proximity to cities (in rural areas) and size of cities (in urban areas). This approach complements the population-based approach of Gollin et al. (2021).[6] However, the city-based approach more explicitly accounts for agglomeration effects[7] since the high population must also be clustered together to be considered a city.

I find that living standards tend to increase monotonically with proximity to cities for rural areas and the city-size for urban areas. The result is robust to whether the consumption index is constructed using Young's (adapted) approach or principal component analysis (PCA). These results are aligned with the urban spillover literature, but have (to my knowledge) never been verified at this scale. Additionally, my findings corroborate the previous emphasis in development economics on the urban-rural divide specifically – this is the largest regional divide. However, other regional inequality can be substantial. I find that on average within-urban inequality is larger than urban-rural education inequality.[8]

My results are also able to partially reconcile the underwhelming panel studies with the cross-sectional studies finding large urban-rural gaps. I find that living

[6] Where quality of life indicators were shown to rise with population quartile.

[7] Agglomeration effects (formalised by Marshall (1890)) are a common explanation for the higher productivity of cities, although identifying the direction of causality is challenging. Recently, the empirical base for these effects has grown (Greenstone et al., 2010; Kline and Moretti, 2014).

[8] When measured using Young's (adapted) approach.

standards between the most well-off rural area and the least well-off urban area are approximately two-thirds of the aggregate urban-rural gap. Consequently, even under the frictions (rather than the sorting) paradigm a migrant may experience less gains than the aggregate cross-sectional consumption gaps would suggest.

I then include informal settlements in the analysis. I focus on informal settlements in the largest city of each country, where they are particularly relevant. I find that common consumption indices, including the adapted Young approach, are inadequate for encapsulating living standards in these areas. Consumption indices often assume away substitution effects to give items in the index a constant weight across the sample. If these indices are consistent with each other, and robust to excluding certain products and product groups, then this assumption is benign. However, the results for slums are not robust to the inclusion/exclusion of government services like health and education.

To explore this further, I decompose the consumption indices into individual items in the categories of housing conditions, asset ownership and children's health/education. I find that conditions in slums tend to be worse in all categories compared to the rest of the largest city. Despite this, on aggregate all measures of living standards are still higher in slums compared to even the most well-off rural area. An exception, for some countries, is that children's disease burden is higher in slums – breaking the monotonicity found in Gollin et al. (2021).

Theoretically, slums could be a bottleneck for migration (in these countries) if households value children's health outcomes sufficiently – and choose not to migrate to urban areas for this reason. However, since the higher disease burden does not generally translate into higher death rates, I argue that it is still unlikely that conditions in slums hinder migration. Instead, the higher disease burden in slums could be due to population density, while better services – or being better off in other regards – equips households to deal with this harm.

These results also contribute to a growing literature that urbanisation in sub-Saharan Africa bucks the historical trend (Gollin et al., 2021): Historically, urban areas had higher mortality and disease, with higher wages to compensate (Riley, 2001; Williamson, 1982; Kesztenbaum and Rosenthal, 2016; Costa and

Kahn, 2006). This paper shows that even slums now tend to have higher living standards than the most well-off rural areas. Although some countries still have a higher disease burden in slums, better quality services may mean this no longer increases mortality rates.

This paper is organised as follows: Section 2 describes the main data sources used for measuring consumption and identifying regions. The sample is largely chosen based on the availability and quality of this data. Section 3 describes the methodology developed in this paper for linking survey data with spatial characteristics and demonstrates the improvement over current best practices. The following two sections then provide detail on the survey and spatial elements to be linked. Section 4 explains how regions were identified, including the creation of a transcontinental map of informal settlements. Section 5 describes how consumption indices were created from household survey data. Finally, section 6 provides the results of the decomposed urban-rural gap, and verifies their robustness. Section 7 concludes.

2 Data, Descriptive Statistics & Sample Choice

Two types of data are needed to measure living standards across decomposed regions. The first is data related to consumption. Following Young (2013) and Gollin et al. (2021), I use the Demographic & Health Survey (DHS) dataset. The second type of data is needed to identify regions. This includes the Global Human Settlements (GHS) Layer (to identify cities) and the Open Buildings dataset (to identify informal settlements). This section describes the sample choice and timing and then provides an overview of the specific data sources.

Sample Choice: The inclusion criteria for the sample is as follows: countries in sub-Saharan Africa that are (i) more than 5 million in population; (ii) not in West Africa (due to constraints in the buildings dataset); and (iii) have the relevant data available, including a geo-coded DHS survey post 2010.[9] Criterion (i) is to restrict the analysis to contexts that tend to be more data-rich, and so are also more comparable to the rest of the literature. It is similar to the inclusion criterion of Gollin et al. (2021) that countries must be of a certain size, except that Gollin et al. (2021) use an area cut-off rather than a population one. Criteria (ii) and (iii) are due to data constraints, including the need for comparable data in the sample across the relevant region/time-frame. The ten countries meeting this criteria can be seen in Table 1. The cumulative population in the sample covers ∼49.68% of the population of sub-Saharan Africa, and a considerably larger proportion of sub-Saharan Africa excluding West Africa (World Bank, 2020).

Timing: The two types of data needed to measure consumption across decomposed regions (consumption and spatial data) are combined using the methodology described in section 3. Importantly, the timing of the datasets is not completely identical, partly since due to the size of these projects they do not occur yearly. Open Buildings uses the most recently available (as of 2021) satellite

[9] Mozambique is excluded from the sample as its consumption index model (discussed later) does not converge, this is assumed to be based on data constraints – not having sufficient DHS clusters overlapping with informal settlements.

data, GHS is an estimate for 2015 and the DHS surveys included in this paper are completed between 2013 and 2020. Therefore, an implicit assumption in combining these sources is some stability in the nature of regions. This is a common implicit assumption. For example, Gollin et al. (2021) also combine population data from a single year with survey data that has slight timing differences (by some years) across countries.

Table 1: Set of Countries in Sample

	Population (million)	**DHS Survey Date**
Angola	32.87	2015-2016
Burundi	11.89	2016-2017
Congo, Dem. Republic	89.56	2013-2014
Ethiopia	114.96	2016
Kenya	53.77	2015
Malawi	31.26	2015-2016
Rwanda	12.95	2019-2020
South Africa	59.73	2016
Tanzania	45.74	2015-2016
Zimbabwe	14.86	2018

Notes: Population statistics from World Bank (2020)

Demographic & Health Survey Data for Measuring Consumption: Following Young (2013) and Gollin et al. (2021), I use the DHS to measure real consumption. The DHS is a large, nationally representative household survey. A valuable feature of the survey is that it is designed similarly across countries so as to provide comparable results (Gollin et al., 2021).

The survey includes an urban-rural variable based on national reporting. This is what Young (2013) used to define regions in his seminal paper measuring the urban-rural consumption gap. Additionally, the households are in geo-coded clusters, allowing (when combined with other data) a more granular understanding of the location of respondents. However, in order to preserve household anonymity, reported locations are displaced. Section 3 explains how the displacement occurs and is addressed.

I use three categories of consumption: Asset (durables) ownership, housing conditions and health/education. These variables largely overlap with those measured in Young (2013), with the exception that (for simplicity) I exclude many variables that Young includes in a 'family time' category. Based on robustness checks done by Young, this should not change the magnitude of results (although it could lead to larger confidence intervals).

Table 2: DHS Real Consumption Means Across Sample

	N	All	Urban	Rural
Health & Education				
All attend school (age 6-14)	126431	0.85	0.91	0.82
All no fever (under age 3)	106411	0.71	0.76	0.69
All no cough (under age 3)	106438	0.67	0.69	0.67
All no diarrhea (under age 3)	106322	0.79	0.81	0.78
All alive (under age 3)	149591	0.76	0.81	0.74
Housing Conditions				
Electricity	233671	0.29	0.64	0.12
Tap drinking water	230584	0.18	0.44	0.06
Flush toilet	223900	0.15	0.40	0.03
Constructed floor	232001	0.42	0.76	0.25
Asset Ownership				
Television	233662	0.25	0.56	0.10
Telephone	233659	0.67	0.87	0.56
Refrigerator	233658	0.14	0.33	0.04
Radio	233679	0.48	0.59	0.42
Motorcycle	233662	0.06	0.07	0.05
Car	233662	0.06	0.13	0.02
Bicycle	233665	0.24	0.18	0.27

Notes: As in Young (2013), all variables are coded as 0 or 1. N is the count of households. Ownership of assets is included as 1 if the household has at least one of the item. Constructed floor includes all floors not made of dirt/sand/dung. Children's health outcomes are coded as 1 if none of the children in the household experienced fever/cough/diarrhea in the past two weeks.

The included variables are chosen for a number of reasons: (i) in developing countries they often form the majority of household spending, with extensive variation allowing a consumption index based on them to be informative; and (ii) they include publicly available amenities, and thus can more adequately reflect overall quality of life (Young, 2013). Table 2 provides descriptive statistics for these variables across the sample. Since the majority of the countries in the sample are low-income, it is not surprising that substantial portions of the population do not have access to each consumption variable. It is also noteworthy that all quality of life indicators, except bicycles, are more prevalent on average in urban areas compared to rural areas. This already reflects the urban-rural divide.

Global Human Settlements Layer (GHSL) for Identifying Cities: The GHSL combines satellite data and population statistics to represent degrees of urbanisation at a $\sim$1km^2 resolution.[10] It was created in order to provide a uniform definition of urbanisation across countries, and is supported/published by the European Commission, Joint Research Centre and GEO Human Planet Initiative. Specifically, I use its definition for cities as either (i) contiguous pixels with a population of at least 1500 per pixel, where pixels are at $\sim$1km^2 resolution; or (ii) build-up density of at least 50% and population of at least 50 000.

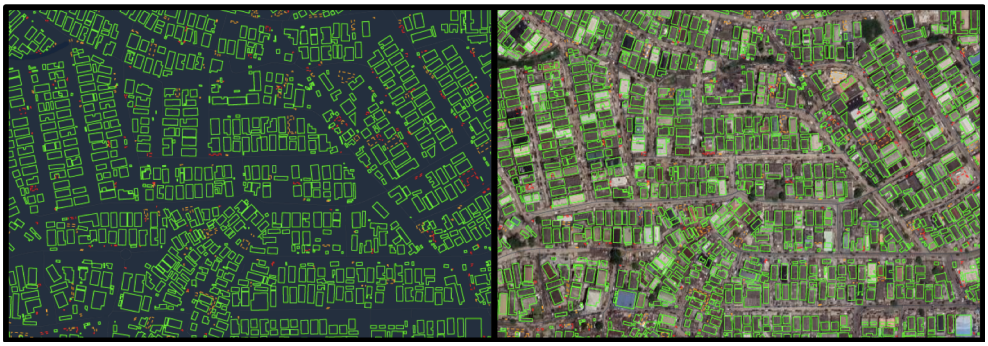
Open Buildings Data for Identifying Slums: Google Research used a machine learning model on satellite imagery (at 50cm resolution) to create a map of buildings across most of Africa, and released the 'Open Buildings' dataset in 2021. This project was at unprecedented scale and resolution, more than doubling the number of buildings known on the African continent. The training dataset included 1.75 million building occurrences that were labelled manually. The final output was the identification of 516 million building outlines on the African continent. Each building is accompanied by a confidence score (Sirko et al., 2021). An illustration of this output can be seen in Figure 1. This data is uniquely appropriate for identifying slums by virtue of the high resolution (50cm), which

[10] The resolution is 30 arc-seconds which is $\sim$1km at the equator.

allows identification of the small buildings that characterise slums.

In addition to this output, the technical report includes evaluation of the model's performance. The report specifies the precision-recall trade-off and investigates how it could vary across geographic locations (both urban-rural differences and between countries, as seen in Figure 2). In this context, recall refers to the proportion of buildings in the area that exist that are identified as buildings (by the model). Precision refers to the proportion of items classified as buildings that are actually buildings. Google Research also provides local confidence score thresholds which can be used to filter buildings to ensure constant precision across geographic area. This is illustrated by the different outline colours in Figure 1.

Figure 1: Example of Open Buildings Data in an Area in Lagos

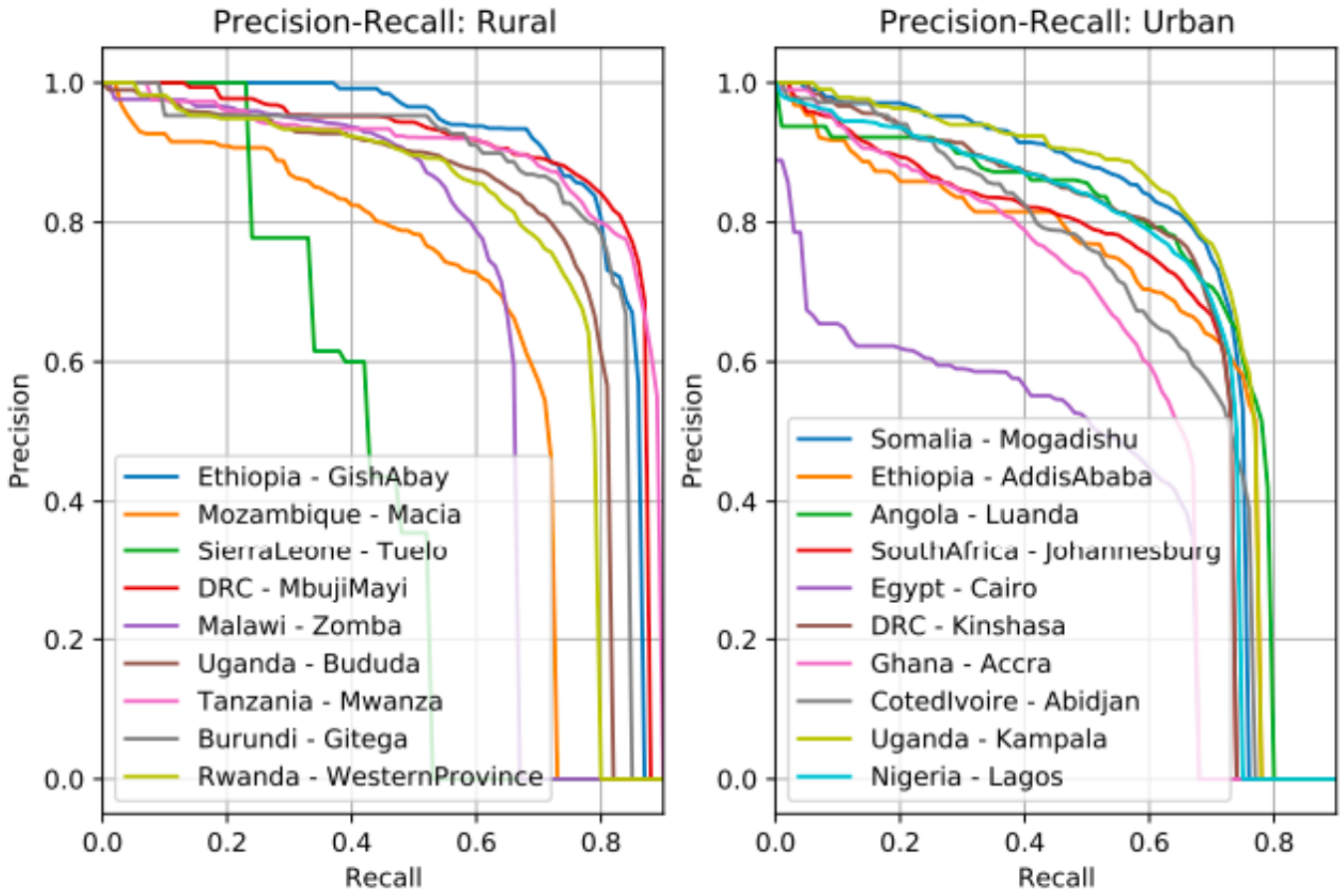


Notes: In the left panel, outlines are buildings identified by the model. The colour represents their confidence score. Green is above 0.7, red is below 0.65 and orange lies between green and red. The right panel is the same data overlaid on satellite imagery.

For reasons explained in section 4,[11] I sometimes use the lowest precision in order to maximise recall. As can be seen in Figure 2, while rural areas have widely varying recall-precision levels, for urban areas (where this data is relevant) most sub-Saharan countries' recall converges (between 70-80%) at low levels of precision. Ghana, however, does not converge and Cote d'Ivoire is the slowest converge. Since these outliers are both in West Africa, I exclude West Africa from my sample in order to have more regional homogeneity across the sample in model performance.

[11] Since slums may disproportionately contain buildings identified with low precision, increasing precision may omit certain informal settlement areas from the buildings data entirely (see section 4).

Figure 2: Evaluation of Precision-Recall in Open Buildings Data Set



Notes: Graphs reproduced in their entirety from the technical report (page 11) of Open Buildings (Sirko et al., 2021). All countries analysed are in sub-Saharan Africa, except Egypt.

3 Connecting DHS to Spatial Data

To measure consumption across decomposed regions, spatial data must be linked to consumption data. For this to be possible, it is necessary to address the displacement of DHS survey locations. I use a more sophisticated probability weighting than is the current best practice. This approach is necessary to decompose regions more granularly (like by informal settlement) than was previously thought possible (Gollin et al., 2021). In this section, I develop an image representing the probability mass of the actual (pre-displacement/true) survey observation. I then explain how this image can then be combined with other images/maps (e.g. population/distance/informal settlement) to provide an accurate probability (if binary) or expected value (if continuous) of that characteristic at the true point. This template is rolled out over hundreds of thousands of DHS observations in under five minutes.[12] Consequently, this methodology is also computationally feasible.

Context: To preserve anonymity, the DHS reports displaced locations of survey clusters instead of reporting actual locations. The maximum displacement is 2km in urban areas and 5km in rural areas (with 1% displaced up to 10km). This section focuses on urban areas, but the analysis could easily be extended to rural areas. In urban areas, the displacement occurs as follows: (i) the displacement angle (θ) is chosen randomly from a uniform distribution between 0 and 360 degrees; (ii) the displacement distance (d) is chosen randomly from a uniform distribution between 0 and 2km. Therefore, the reported location (l_r) differs from the actual location (l_a) – although l_r will always be within a 2km buffer of l_a.

This displacement makes identifying the neighbourhood of the actual location challenging due to potential classification error.[13] The DHS guidelines[14] suggest addressing this by calculating the proportion of the area of the 2km buffer that overlaps with the neighbourhood and using that to proxy probability. These

[12] On the publicly available platform Earth Engine.

[13] Particularly for points close to the boundaries of neighbourhoods.

[14] These are considered DHS's guidelines in that they are published by the DHS, in the Perez-Haydrich et al. (2013) paper 'Guidelines on the Use of DHS GPS Data.

guidelines are widely followed (Gollin et al., 2021). However, this loses information by assuming points are uniformly distributed within that buffer (which they are not: the displacement process leads to points concentrated close to the centre of the buffer, as illustrated later). So instead, I approximate $Pr(l_a|l_r)$ – the probability mass function of l_a given l_r – more accurately by fully taking into account the displacement procedure.

I am not the first to recognise that there are feasible improvements to the DHS guidelines. Warren et al. (2016) show that more fully accounting for the displacement procedure outperforms the naive approach for point-in-polygon analysis. I adopt a similar approach, but my approach is simpler and easier to implement because it takes advantage of the symmetry between $Pr(l_r|l_a)$ and $Pr(l_a|l_r)$, described below.

Theory for estimating $\boldsymbol{Pr(l_a|l_r)}$: Recognising the symmetry between $Pr(l_r|l_a)$ and $Pr(l_a|l_r)$ allows $Pr(l_a|l_r)$ to be estimated as straightforwardly as $Pr(l_r|l_a)$ can be. The argument for this is as follows:

Premise 1: *If the actual location (l_a) were known, it would be straightforward to calculate $Pr(l_r|l_a)$ given the displacement procedure described previously.*

The distribution of $l_r|l_a$ is uniquely set by the distribution of d and θ. One method of approximating $Pr(l_r|l_a)$ would be to use a Monte Carlo simulation to generate a multitude of points according to the displacement procedure. The proportion of points in each area would then represent the probability that l_r lies in that area. Perhaps a more efficient alternative to the Monte Carlo simulation would be to find a closed form solution. The density of the point l_r depends on the joint density of d and the location on the circumference of the circle:[15] $f(d, circle_l) = 1/(4\pi d)$.[16] The probability of l_r lying in an area could then be found by integration.

[15] The circle's circumference is $2\pi d$, the location on the circle's circumference ($circle_l$) is determined by the angle θ.

[16] $f(d, circle_l) = f(d)f(circle_l) = (1/2)(1/(2\pi d)) = 1/(4\pi d)$, since (i) d and θ are independent; (ii) The circle's circumference is $2\pi d$, the location on the circle's circumference ($circle_l$) is determined by the angle θ; and (iii) $d \overset{iid}{\sim} U[0, 2]$ and $\theta \overset{iid}{\sim} U[0°, 360°)$.

Premise 2: *$l_a|l_r$ and $l_r|l_a$ have the same distribution, due to symmetry.*

The distance (d) between l_a and l_r is by definition identical to the distance between l_r and l_a, where $d \overset{iid}{\sim} U[0, 2]$. The angle ($\theta$) between l_a and l_r has the same distribution as the angle (α) between l_r and l_a. As $\alpha = 180° + \theta$ and $\theta \overset{iid}{\sim} U[0°, 360°)$, due to the linear transformation of a uniformly distributed variable, $\alpha \overset{iid}{\sim} U[180°, -180°)$. Therefore, both θ and α are uniformly distributed around the angles of a complete circle, and $\alpha \overset{iid}{\sim} U[0°, 360°)$.

Conclusion: *One can estimate $Pr(l_a|l_r)$ in the same straightforward manner as $Pr(l_r|l_a)$.*

Implementation: Based on the above theory, I develop an image representing the probability mass of the actual (pre-displacement/true) survey observation ($Pr(l_a|l_r)$). I use the Monte Carlo simulation (described previously) to obtain $Pr(l_a|l_r)$. While the closed-form solution is viable, I choose the simulation due to its ease of implementation.[17] By generating enough points, it should closely approximate the closed-form solution. Additionally, achieving more granular accuracy than 500m^2 resolution is not necessary given the resolution of other spatial data used in this paper.

I simulate 100 000 pairs - indexed by j - of (d, θ) according to their respective uniform distributions. I calculate the [x,y] coordinates of each simulated $\hat{l}_{a_j}$, assuming l_r lies at the origin of a grid (at coordinates [0,0]). Using basic trigonometry, this implies that $[x_j, y_j] = [d_j \cos\theta_j, d_j \sin\theta_j]$. The x coordinate represents the horizontal displacement (in km) from l_r, and the y coordinate represents the vertical displacement (in km) from l_r.

I translate this information into an image template (pixel representation) on a map for each country. First, I select the centroid of the country – this will act as l_r^m. The m superscript indicates that the point is now on the map.[18] Then, I translate the simulated $\hat{l_a}$ coordinates from a km representation to coordinates

[17] On the platform Earth Engine, it is easier to reduce geometries (points) to an image (the probability template, discussed later), than to integrate across each pixel in an image as the closed-form would require.

[18] Rather than at the origin of a km grid.

in longitude and latitude (relative to l_r^m). I obtain the local conversion rate[19] by adding 0.01 degrees to both the longitude and latitude coordinates of l_r^m and finding the resultant kilometre changes (see Figure 3).

Figure 3: Example of l_r^m and Local Kilometre to Coordinate Conversion

Notes: The point is an example of l_r^m – the assumed reported location on the map (in this case, it is the centroid of Zambia). Both lines represent a 0.01° displacement in the coordinates of l_r^m. The green line (0.01° in latitude) is 1.094km, and the blue line (0.01° in longitude) is 1.106km. The conversion rates (c_x and c_y) are calculated from the ratios between degrees and kilometers, namely $c = \Delta degrees / \Delta kilometers$.

I then apply that conversion to each $\hat{l}_{a_j}$.[20] Finally, I create an image of resolution 500m^2, where each pixel represents the number of points in that area. I divide this image by j to create the final image (which I will refer to as the Probability Mass Template), where each pixel approximates the mass of $l_a^m | l_r^m$ in that area.

An example Probability Mass Template is displayed in Figure 4. As expected, the mass monotonically decreases from the centre of the circle and reaches 0 after a 2km circular buffer. This pattern coincides with the closed form solution described previously. I also verify that the sum over the pixel values is 1. There is slight asymmetry, which is the consequence of not choosing l_r^m to be in the intersection or centre of grid-lines.[21]

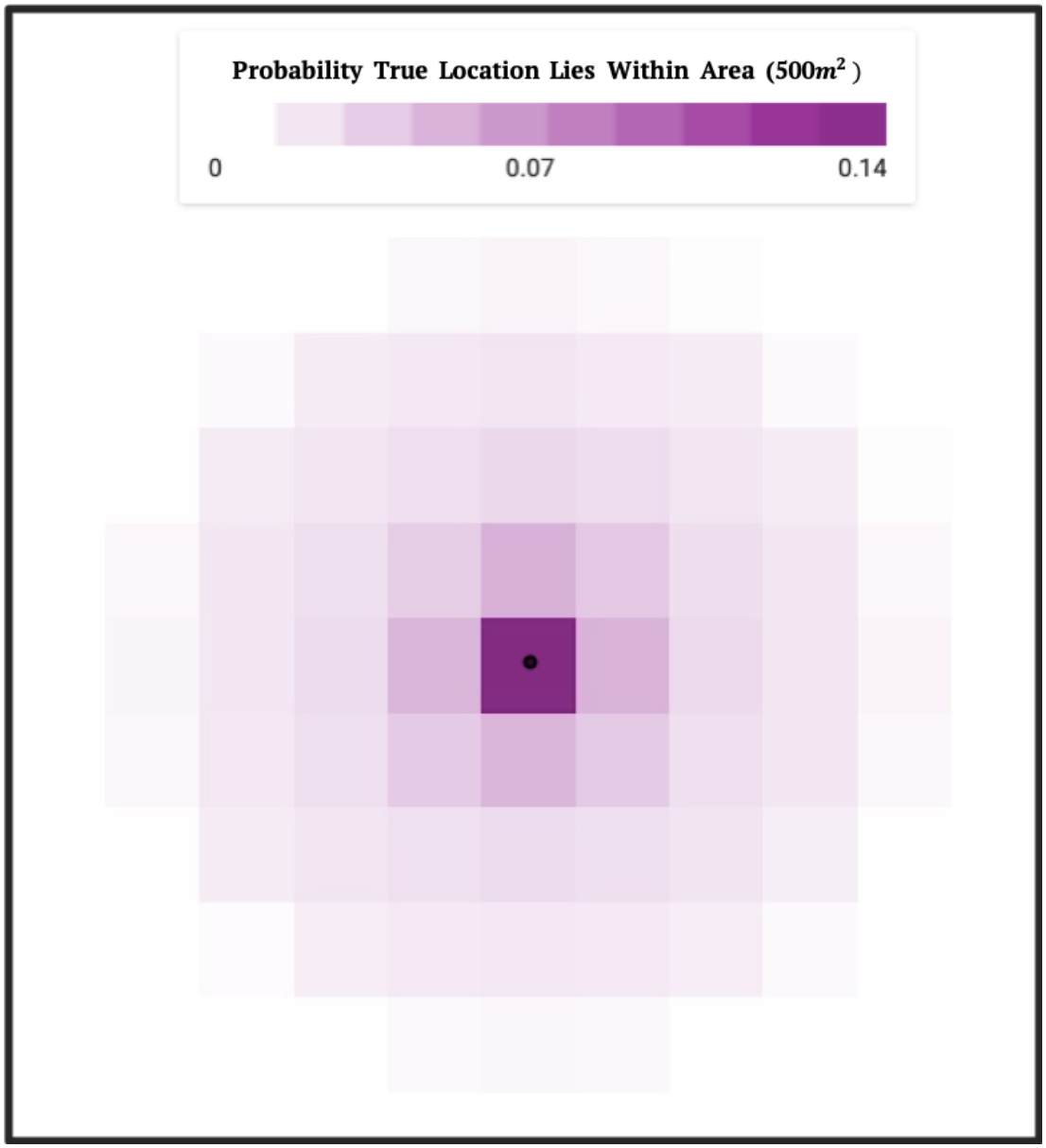
As can be clearly observed from Figure 4, the actual location is considerably more likely to be close to the reported location than further away from it. The use of this template is therefore demonstrably better than assuming a uniform distribution, as is the current best practice.

[19] The map uses a WGS84 projection, which implies that a unit km change may translate into different degrees of longitude and latitude depending on how close a point is to the equator.

[20] Formally: $l_r^m = [\text{longitude}_r, \text{latitude}_r]$ and $\hat{l_{a_j}^m} = [\text{longitude}_r + x_j * c_x, \text{latitude}_r + y_j * c_y]$, where c_x and c_y represent the conversion ratios (between kilometres and longitude/latitude respectively) identified via the process described previously and illustrated in Figure 3.

[21] Resolution grid-lines allow images/pixels to be overlaid exactly.

Figure 4: Probability Mass Template



Notes: The image is a Probability Mass Template, each $500m^2$ pixel represents the probability of the actual location given the reported location. It is created via the Monte Carlo simulation described in this section. The (black) point is the l_r^m (assumed reported location on the map) chosen in Ethiopia. As can be clearly observed from this image, the actual location is considerably more likely to be close to the reported location than further away from it. The use of this template is therefore demonstrably better than assuming a uniform distribution, as is the current best practice.

Using the Probability Mass Template: I then combine the Probability Mass Template with other spatial data. For example, to get the probability that a household lives in an informal settlement, I do the following: (i) move the Probability Mass Template to surround a DHS cluster;[22] (ii) multiply the Probability Mass Template by the map of informal settlement (which is one where slums are and zero elsewhere); and (iii) sum the pixels of the new image to get the probability that the DHS cluster is in a slum.

I complete the above steps for each DHS cluster simultaneously (which saves time compared to completing them iteratively). Using Earth Engine, this template can be rolled out over hundreds of thousands of DHS observations in under five minutes. Notably, this process could also be applied with continuous (rather than binary) maps (e.g. population density), with the output being interpreted as an expected value rather than a probability.

Strengths and Limitations: Using the Probability Mass Template more accurately estimates $Pr(l_a|l_r)$ than following the DHS Guidelines, as all information in the displacement procedure is used. The limitations of my approach are due to simplifications to reduce computing power and achieve scale.

The Probability Mass Template is a WGS84 projection, while the DHS displacement is in kilometres – this could cause distortions if the template was used over large distances. The distortion would arise because the latitude to km conversion depends on distance to the equator. To reduce this distortion, a new template (which takes into account local conversion) is created for each country.

An additional limitation is that DHS displacement ensures clusters do not change administrative boundaries. Therefore, a more accurate Monte Carlo simulation would, for some DHS locations, not simulate points in a perfect circle. My approach is more accurate when displacement does not often overlap with administrative boundaries. However, both my approach and the DHS guidelines have this limitation.

[22] This movement can create slight inaccuracies if the DHS cluster is located differently in the central pixel. The maximum size of the error is the size of the central pixel.

4 Identifying Regions

This section describes how regions are identified, this identification will be critical for evaluating living standards in following sections. Regions are decomposed as follows: (i) by proximity to cities in rural areas; (ii) by size of city in urban areas; and (iii) by residence in an informal settlement in the largest city of each country. The first subsection focuses on measuring the distance to and size of cities. Least-cost maps are created representing distance to cities in ten sub-Saharan countries. The second subsection concentrates on identifying informal settlements – which involves creating the first map of informal settlements across sub-Saharan Africa. Due to potential heterogeneity within slums (based on how planned/formal they are), I also develop a tool to proxy for planning by measuring how regularly (on orthogonal axes) buildings are laid out. The final subsection describes this tool. The probability weighting I developed in section 3 is used extensively in this section.

4.1 Using Cities to Decompose Urban/Rural Areas

I divide rural areas into subregions based on proximity to cities, where a city is defined as in the GHSL (see section 2). I use a least-cost formula to create a map of each country, where each pixel represents the shortest distance in meters to a city.[23] I assume that borders cannot be crossed and implement this constraint by specifying a cost sufficiently high to crossing the border, such that it is never optimal for the least-cost formula to include breaching it. This constraint is particularly influential when countries are irregularly shaped, such that a direct path between two points within a country could involve crossing the border multiple times. For simplification purposes, I do not increase the cost for natural obstacles (e.g. rivers/mountains).

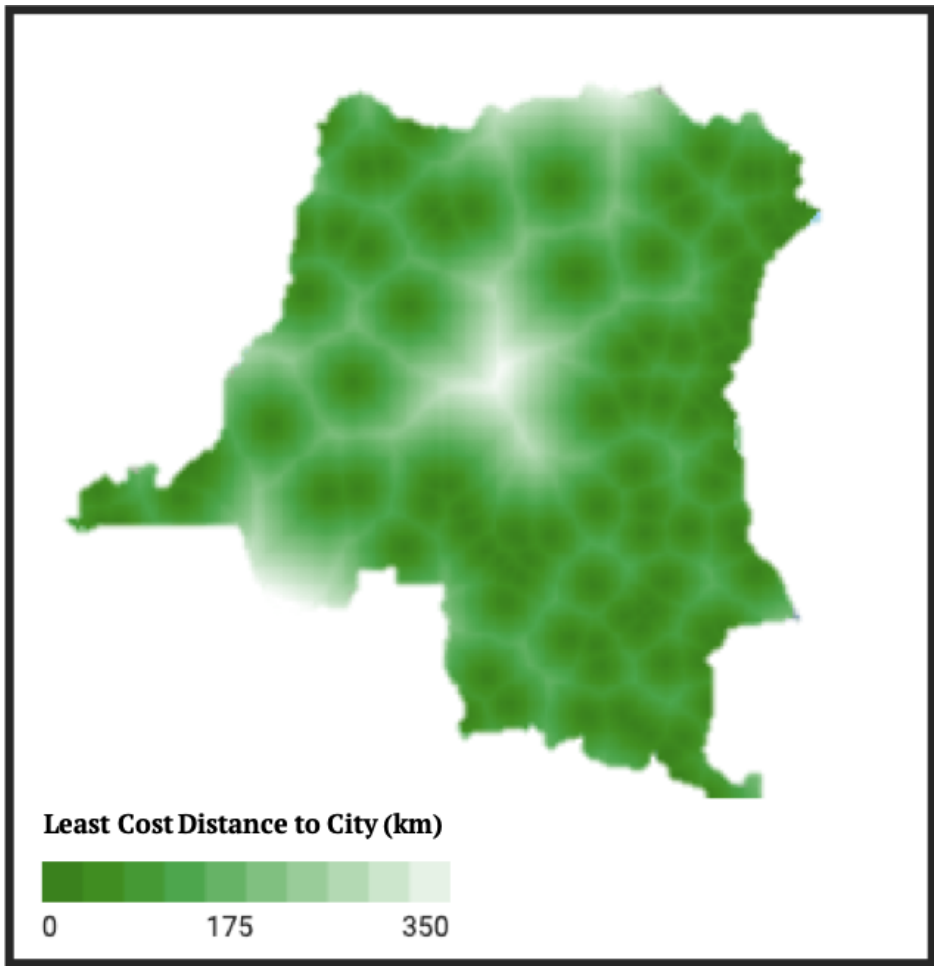
An example of the final output can be seen in Figure 5, which illustrates a relative dearth of cities in the central part of the Democratic Republic of the Congo.

I then create quartiles of rural DHS clusters based on distance to cities - using the DHS's classification of 'rural'. I do not use probability-weighting because (i) the

[23] The least-cost formula calculates the least cost path between a city and each pixel on the map. The cost of traversing each pixel, within the country, is the size of that pixel in meters.

distances in rural areas tend to be large relative to DHS displacement; and (ii) I am dividing into quartiles where the displacement would only matter for clusters on the boundary (of which there are unlikely to be many). Consequently, the displacement is likely to have a negligible impact on the quartile categorisation (Perez-Haydrich et al., 2013).

Figure 5: Least-Cost Distance to City in Democratic Republic of Congo



I also use a city-based approach to classify urban areas. I divide urban areas into three regions: The largest city (by area), other cities, and DHS urban areas that are not in any city. Since cities may only be a few square kilometres in size, I use probability weighting to categorise DHS clusters.

The category of the urban DHS cluster is then the area they are most likely to be in (with a probability of over 50%). To obtain this probability, I use the Probability Mass Templates I developed in the previous section. I also use this method when identifying households most likely to be living in informal settlements.

4.2 Identifying Informal Settlements

I use the Open Buildings dataset to classify areas within cities across the African continent as 'slums' or 'not slums'. To my knowledge, this is the first international map of slums in Africa. I identify informal settlements using their geographic footprint, loosely following Michaels et al. (2021) characterisation of slums as collections of small, dense buildings in urban areas. This definition enables the identification of slums despite not having data on the height of buildings, since the small base area of slum buildings allows the assumption that the base could not support a height of more than one story.

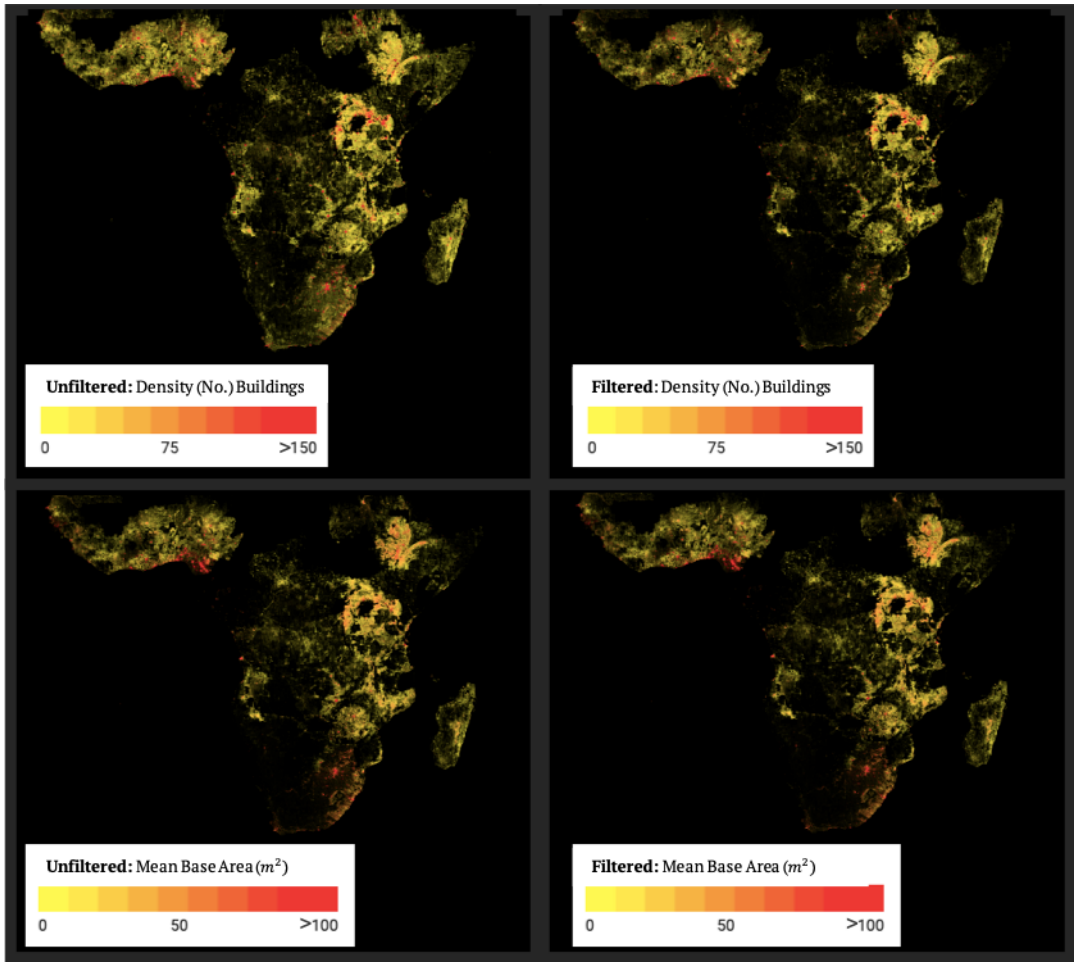
I begin by creating maps across sub-Saharan Africa representing the density and mean base area of buildings (unfiltered and filtered for 80% precision) at $500m^2$ resolution. I do this by reducing the geometries of the Open Buildings dataset to an image based on their characteristics in that area/pixel. This is computationally expensive, so I use parallelisation (with cloud computing) to reduce waiting time. The result is four datasets spanning sub-Saharan Africa, with potential uses beyond this paper (see Figure 6) – these uses will be discussed further in section 7. These maps allow areas with small, dense buildings to be identified.

I then identify informal settlements as follows: Within cities (classified as in the previous subsection), informal settlements are those regions with a mean (unfiltered) building base area below $50m^2$, where the mean is taken for each $500m^2$. For the analysis in this paper, I focus on slums in the largest city in the specific (ten) sub-Saharan African countries in the sample, as slums may be most relevant as a bottleneck to migration in the largest city (which is often also the capital).

An alternative method to identifying informal settlements would be to filter by high building density (rather than low building mean base area). These metrics are related in that very high building density may imply a lower mean base area, simply because many buildings must fit into the specified space. Figure A1 in the appendix confirms building mean and building density (in the largest cities) are negatively correlated across the sample. However, using building density to identify informal settlements would not be ideal in this context because this measure may

be particularly sensitive to recall (which I am unable to standardise).[24]

Figure 6: Building Mean Base Area and Density Maps Across Sub-Saharan Africa



Notes: These are the datasets created in this paper, they span sub-Saharan Africa (with the exception of Cameroon and South Sudan). Building mean base area is in m^2 and density is in no. of buildings per $500m^2$. The left panel are maps based off unfiltered buildings. The right panel are maps based off buildings filtered to 80% precision. Black represents no buildings in the dataset. These maps are created by reducing buildings in Open Buildings dataset to an image based on characteristics. They can be made publicly available on Earth Engine.

This sensitivity is reflected in Figure 7, which shows that (compared to mean base area) density measures are less robust to the confidence filters chosen. While the correlation between filtered and unfiltered mean base area is 0.94 at DHS locations, it drops to 0.72 for filtered and unfiltered building density. These correlations are obtained by using the probability weighting I developed in section

[24] See section 2 for extensive discussion of the recall-precision trade-off in this dataset.

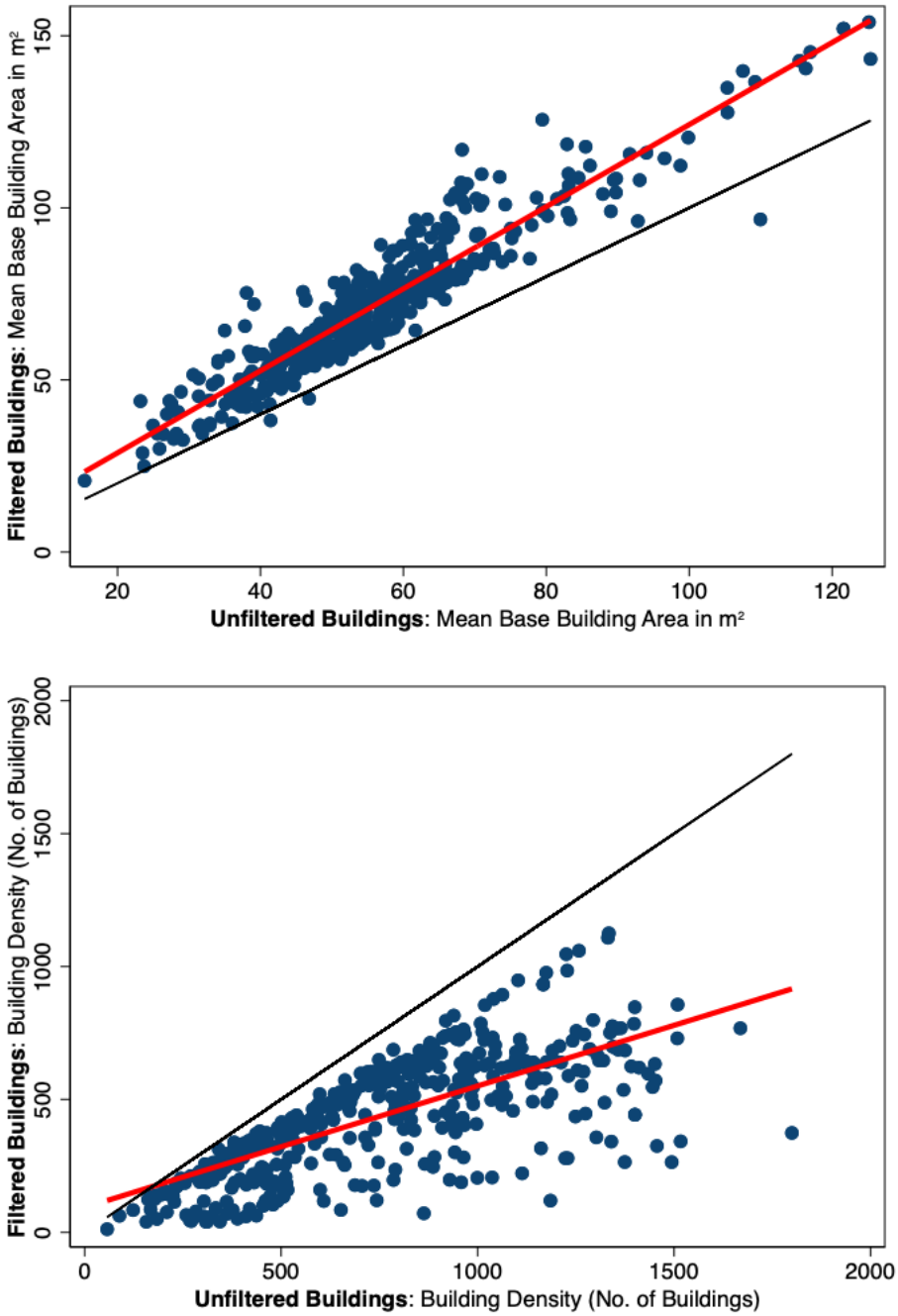
3 (combining the Probability Mass Templates at DHS locations with the maps in Figure 6). These patterns appears to apply on a grander scale than just the largest city. Visually, one can also see that in Figure 6 (which are the datasets created in this paper covering most of sub-Saharan Africa) the filtered compared to unfiltered mean base area map appears more similar than the same comparison on density maps.

Consequently, using the mean base area measurement rather than density is more robust to the modeler's choices. Further, the mean base area measure is largely geographically consistent – given that I can only standardise precision (not recall), and the mean base area measure appears relatively insensitive to recall.[25]

To identify slums by mean base area, I use the unfiltered mean base area map rather than the one filtered for uniform precision. Theoretically, the trade-off between filtering or not filtering is as follows: Not filtering allows more false positives (and potentially geographical inconsistency if the percentage of false positives varies by region), but it also mitigates the risk that informal settlements (which may have buildings with lower confidence thresholds) are omitted entirely. Figure 7 confirms that smaller buildings get omitted when thresholds are applied, as the unfiltered mean base measure tends to be lower than the filtered one. In this case, using lower precision to maximise recall would ensure informal settlements' inclusion. Empirically, as discussed previously and illustrated in Figure 7, the difference between using either filtered or non-filtered buildings appears negligible. Once again, I categorise DHS clusters as being in slums if the probability is over 50%.

[25] In the sense of being insensitive to the recall-precision trade-off (Figure 7).

Figure 7: Correlation Between Unfiltered and Filtered Measures in Largest Cities Across Sample



Notes: Each figure shows the correlation between (i) an unfiltered building measure and (ii) the same measure but with buildings filtered to achieve uniform (80%) precision. The red line is the line of best fit, and the black line is the 45° line. The measures are taken at DHS clusters across the entire sample (the largest cities of 10 countries), and probability weighting (combining the maps in Figure 6 with the Probability Mass Template, both at $\sim 500m^2$ *resolution) is used to get expected mean and density. The mean measurements (top panel) have a correlation of 0.94, and the density measurements (bottom panel) have a correlation of 0.72.*

4.3 Within-Slum Inequality: Creating a Proxy for Planning

Within informal settlements, I make a final distinction based on how regularly laid out the buildings are (and divide informal settlements into two groups based on this measure). Similarly to Michaels et al. (2021), this measure proxies for more formal/planned areas in slums, which may coincide with quality. Consequently, not only is this paper the first to evaluate living standards in slums at this scale, it also attempts to quantify within-slum inequality.

In the Michaels et al. (2020) paper, regular-layout was measured in slums of a single city. A bounding box was manually drawn for each building, and the angle of each bounding box from its neighbour was measured. Since my dataset contains millions of buildings, this approach would be infeasible and automation is necessary. Consequently, I create a tool that calculates the standard deviation of the angles of building walls from two orthogonal axes, where the axes are chosen to minimise that deviation.[26] This is done in the following way for buildings in a specific area:

I first choose an initial (arbitrary) set of orthogonal axes[27] and measure the angles of walls (of buildings in the area) relative to these axes. To measure the deviation of the angles from the initial axes, I apply modulus 90. Effectively, this places all angles between 0° and 90°.

I then calculate variance. This requires a mean measurement. A simple arithmetic mean of the transformed angles would be misleading, especially if angles were close to the extremes of 0° or 90°. For example, 5° and 85° would be treated as if they were 80° apart, meanwhile actually, since 5° is equivalent to 95°in terms of deviation from orthogonal axes, they are much closer (10°apart). This issue is resolved by using the circular mean. The (N) transformed angles are placed on a circle (multiplied by four). The circular mean is calculated as follows, where α_{cj} is a particular angle (j) on the circle (c) and μ_c is the circular mean:

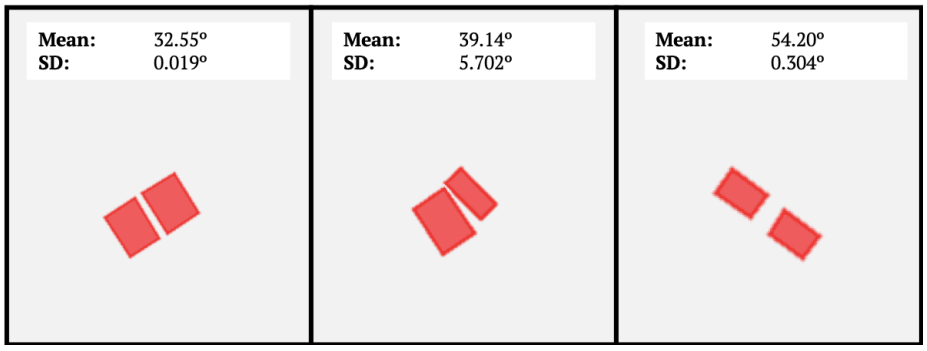
$$\mu_c = \text{atan2}\left(\frac{1}{N}\sum_{j=1}^{N}\sin\alpha_{cj},\ \frac{1}{N}\sum_{j=1}^{N}\cos\alpha_{cj}\right), \tag{1}$$

[26] The axes are the circular mean and the line perpendicular to it, as will be described later.

[27] For simplicity, I choose lines that correspond to longitude and latitude.

I re-normalise the circular mean and angles (divide by 4).[28] The standard deviation is then obtained using the conventional formula.[29] Intuitively, to achieve zero standard deviation ('perfect regularity'), all walls in an area would have to lie precisely on a set of orthogonal axes (the circular mean or the line perpendicular to it). The output of the tool at a small scale is exhibited in Figure 8.

Figure 8: Illustration of Regular Layout Tool



Notes: Red represents buildings. The tool's output is (circular) mean and standard deviation (of wall angles from orthogonal axes). The buildings in the left and right panel have a standard deviation of less than 1°: They show aligned buildings. The middle panel shows buildings less regularly laid out, with a standard deviation of 5°, but on a grid close to the left panel (with a similar circular mean).

This tool aims to get a representative measure of how regularly laid out buildings are in informal settlements and so proxy for more formal/planned areas. However, there are two limitations to using this measure. First, it is only appropriate when there is high building density. This is because even planned/formal buildings need not be orthogonal/parallel to each other in more spacious settings, as space is not a constraint. This should not be a problem as I intend to only use this tool in slums. Second, this variance measure should not be used on too large an area. For example, if two neighbourhoods are both regularly laid out but on different grids, the variance of the entire area will be relatively high. It would be more reflective of regular layout to take the variance of each neighbourhood and average that.

With these limitations in mind, I scale the tool over the map of slums obtained in the previous subsection. I sample uniformly from slums, such that each pixel

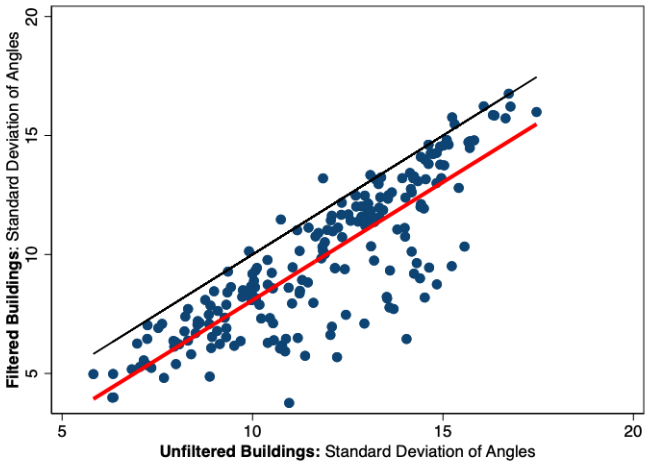
[28] Placing the angles between 0° and 90° again.

[29] $\sigma = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\mu - \alpha_j)^2}$, where μ and α_j refer to the renormalised values, and σ is the standard deviation.

(of 500m^2 resolution) contains over 15 points in expectation. I consider each pixel, rather than each slum, because areas within slums may be heterogeneous.[30] I then apply the regularity tool to each 100m^2 area centred on the randomly chosen points. I choose the size 100m^2 to balance being representative of the building layout around the point while still being relatively small (to avoid containing neighbourhoods on different grids). Since the measure is sensitive to outliers,[31] I take the median of the standard deviations as the value for that pixel.[32]

Finally, I assign each DHS cluster that overlaps with an informal settlement a probability-weighted estimate of how regularly laid out that informal settlement is. I do this by the same procedure described in section 3, except I also re-normalise the estimate (by dividing by the probability of living in a slum) since the standard deviation map is masked everywhere except in informal settlements. The buildings data is not filtered for precision. The final standard deviation measure is reasonably robust to this choice, with a correlation of 0.84 between the unfiltered measure and the measure to obtain 80% precision (Figure 9).

Figure 9: Correlation Between Unfiltered and Filtered Standard Deviation



Notes: Within slums in largest cities, correlation between standard deviation in degrees (of angles from orthogonal axes) when buildings unfiltered and filtered to uniform precision. The red line is the line of best fit, and the black line is the 45° line. Measures probability-weighted and re-normalised. The correlation between the measures is 0.84.

[30] For example, certain areas within slums may have been renovated, while other parts remain informal. Considering pixels enables the distinction between areas to be more accurate.

[31] Outliers in the form of neighbourhoods on different grids, or less dense areas contained in the pixel where space is less of a constraint.

[32] I also filter to include only 100m^2 areas with more than two buildings, this avoids outliers.

5 Creating Consumption Indices

This section focuses on measuring consumption gaps. This is necessary to compare living standards across regions. Ideally, it would be possible to measure overall real consumption via a single representative index. In data-rich environments, this is achieved through deflating nominal consumption expenditure by the price of goods. Unfortunately, this is not possible in sub-Saharan Africa at scale due to data constraints (Lagakos, 2020). Consequently, as discussed in this section, I create a consumption index from DHS consumption indicators (described in detail in section 2). I improve upon the work of the seminal paper measuring the urban rural consumption gap (Young, 2013), by implementing Young's approach as a modification of a simple Item Response Theory (IRT) model. My approach has the advantage of being more transparent, informative[33] and convenient to implement than the original methodology.

Context: Young (2013) pioneers a method, using maximum likelihood, for inferring the urban-rural divide from consumption data. A common interpretation of the method is that it exploits the relationship between education and both (i) product demand equations; and (ii) real income, to infer real consumption overall via the resulting implied Engel curves (Lagakos, 2020; Young, 2012). I show that it can be understood as an extension of Item Response Theory (IRT).

IRT is popular in fields such as psychology and health sciences (Johnson, 2007). While Young does not mention the IRT literature explicitly, his approach has many similar qualities in that it exploits the co-movement across product consumption to identify latent (unobserved) aggregate real consumption. Consequently, I implement Young's approach by adapting a simple IRT model.

Using an IRT framework is useful in several ways. First, it can be implemented easily on Stata with a single line of code, as opposed to the hundreds of lines of code accompanying Young's original paper (Young, 2012).[34] Second, it increases the

[33] The confidence intervals of country measurements can also be captured under my approach. These are unreported in Young's paper.

[34] Young's (2013) paper is not released with code, but a closely related previous paper (Young, 2012) provides code related to the technique.

transparency of Young's approach. A significant disadvantage of Young's approach has been its lack of transparency compared to using expenditure and price data (Lagakos, 2020). Framing the approach as an extension of IRT partially alleviates this by making the log-likelihood function explicit, thereby exposing the model's underlying assumptions. Additionally, the IRT method is easily able to output confidence intervals unlike the original approach. Third, using the IRT framework is useful in that it allows Young's approach to be connected to the broader literature on the creation of consumption indices. For example, IRT has been benchmarked as a quality-of-life index against its main competitor (PCA) and found to perform similarly (Filmer and Scott, 2012).

In this paper, I implement an IRT model (based on Young's approach) on the most recent DHS surveys in sub-Saharan Africa. This provides a measure of the urban-rural gap across countries. I then modify the approach to decompose the urban-rural gap further. I also compare these results to results obtained via PCA as a robustness check. Ultimately, this section describes the methods used for creating consumption indices from household survey data and the circumstances in which the assumptions underlying these methods are appropriate.

Item Response Theory as a Consumption Index: Item Response Theory uses maximum likelihood estimation to identify a latent variable. The co-movement across items allows identification of the latent variable which is assumed to be driving that co-movement. For example, an archetypal use of an IRT model is to estimate (unobserved) ability from several tests (the 'items') attempting to measure that ability.

In this context, the latent variable is the household's logged real consumption. Young (2012) assumes the logarithm of real consumption is proportional to the logarithm of real income through the following relationship:

$$\ln C_h = \alpha + \ln Y_h \tag{2}$$

This is equivalent to assuming a constant marginal propensity to save across households and negligible autonomous consumption. The 'items' (in this simple

model) are a household's (binary) consumption of specific products, which are assumed to be partly driven by logged real income (through logged real consumption, aggregated over all products).

To estimate an index for logged real consumption, assumptions are required on the distribution of both the latent variable (l) and items ($\boldsymbol{Q}$). The likelihood of the simple IRT model, in this case, is as follows:

$$\begin{aligned} L(\boldsymbol{\theta}) &= \int_{\mathbb{R}} f(\boldsymbol{Q}|l,\boldsymbol{\theta},\boldsymbol{X})\ \phi(l|\mu_l,\sigma_l^2)\ \mathrm{d}l \\ l &= \ln C \end{aligned} \tag{3}$$

where $\mathbb{R}$ represents all real numbers, ϕ is the probability density function of the normal distribution, $\ln C$ is the (uni-dimensional) latent variable (logged real consumption) which is assumed normally distributed across households with mean μ_l and variance σ_l^2. Specifying μ_l and σ_l^2 will set the scale and location of the (unobserved) consumption index, and the consequences of this will be discussed later in the section. $\boldsymbol{Q}$ is a vector of the observed product choices (e.g. $q_{hj} = 1$ if household h has product j, and 0 if not). $\boldsymbol{X}$ is a vector of observed household demographics that may have differing effects on product consumption depending on the product. $\boldsymbol{\theta}$ is a vector of parameters, including the coefficients on demographics and loading factors for the latent variable. $f(\boldsymbol{Q}|l,\boldsymbol{\theta},\boldsymbol{X})$ is the conditional joint probability mass function of households' consumption bundles. Formally:

$$f(\boldsymbol{Q}|l,\boldsymbol{\theta},\boldsymbol{X}) = \prod_{h=1}^{n} \prod_{j=1}^{p} f(q_{hj}|l,\boldsymbol{\theta},\boldsymbol{X}), \tag{4}$$

where n is the number of households and p is the number of products. Importantly, this assumes that q_{hj} (given l, $\boldsymbol{\theta}$ and $\boldsymbol{X}$) are independent and identically distributed. This is a restrictive assumption, as will be discussed shortly. Following Young, a random effects model is assumed such that:

$$\begin{aligned} f(q_{hj}|l,\boldsymbol{\theta},\boldsymbol{X}) &= \ell(z_{hj}) \\ z_{hj} &= \delta_j + \boldsymbol{\theta}_{1j}\boldsymbol{X}_h + \theta_{2j} l_h \end{aligned} \tag{5}$$

where ℓ represents the probability density function of a standard logistic

distribution, and δ_j is the product-specific intercept.

Young's Approach: I will briefly describe Young's approach, before illustrating that it can be implemented as a modified IRT model. It will now be helpful to also specify the subscript g to indicate a household's group (for example, rural/urban). As discussed previously, Young (2012) justifies logged real consumption as a latent variable by assuming the following relationship:

$$\ln C_{hg} = \alpha_g + \ln Y_{hg} \tag{6}$$

Young (2013) also assumes that logged real income can be decomposed as follows:

$$\ln Y_{hg} = \ln Y_g + \beta E_{hg} + u_{hg} \tag{7}$$

where $\ln Y_g$ is logged average real income over households (net of education) for a specific group, E_{hg} is the mean of a household's education (of adults), β indicates returns to education, and u_{hg} is household random effects concerning real income. This essentially assumes that real income inequality (within a country) is driven by education differences, group differences, and household random effects. Substituting (7) into (6) implies:

$$\ln C_{hg} = \ln C_g + \beta E_{hg} + u_{hg} \tag{8}$$

where $\ln C_g = \alpha_g + \ln Y_g$ and represents logged average real consumption over households (net of education) for a specific group. E_{hg} is observed, and u_{hg} is assumed normally distributed, such that the latent variable is normally distributed. (8) can be re-written to begin isolating the urban-rural gap:

$$\ln C_{hg} = \ln C_R + (\ln C_U - \ln C_R)d_U + \beta E_{hg} + u_{hg} \tag{9}$$

where d_U is a dummy variable equal to 1 in urban (U) areas and 0 in rural (R) areas. Finally, (9) can be rewritten as:

$$\ln C_{hg} = \ln C_R + \delta\beta d_U + \beta E_{hg} + u_{hg} \tag{10}$$

where $\delta\beta = \ln C_U - \ln C_R$ and δ is the urban-rural consumption gap (net of education) in 'educational equivalent units' (Young, 2013). For example, if $\delta = c$, where c is a constant, the expected (based solely on their observables) change in a household's logged real consumption ($\ln C_{hg}$) from baseline if they lived in an urban area rather than a rural one would be the same as increasing education by c. Young then assumes that, within a country, real demand for product j by household h in group g is given by:

$$\ln Q_{ph_g} = \alpha_j + \boldsymbol{\theta}_{1j} X_{h_g} + \theta_{2j} \ln C_{h_g} + e_{jhg} \tag{11}$$

where $\ln Q_{phg}$ is either the log of real demand or a 'related measure' like a probability index, e_{jhg} is logistically distributed and the other variables are as previously defined. Young substitutes (10) into (11) and then estimates all product equations simultaneously. The code used to achieve this is not released with the Young (2013) paper, but related code can be found in a Young (2012) paper on a similar topic. However, the implementation requires hundreds of lines of code. By contrast, an IRT model can be implemented in a single line, and also outputs confidence intervals which Young (2013) does not report.

Combining Young's Approach and IRT: It can now be seen that the simultaneous estimation of product equations can be done via IRT, due to the equivalence between (5) in the IRT model and (11) in Young's model. Substituting (10), which is justified by Young's approach, into (5) from the IRT model (now including g subscripts) and rearranging gives:

$$z_{hgj} = [\delta_j + \theta_{2j} \ln C_R] + \boldsymbol{\theta}_{1j} X_{hg} + \theta_{2j}(\delta\beta d_U + \beta E_{hg} + u_{hg}) \tag{12}$$

θ_{2j} can now be interpreted as what Young (2013) calls a 'quasi-Engel' curve. This is because it indirectly references the relationship between real income (see (6)) and a product consumption index in the form of the probability index in a logistic regression (see (5)). Equation (12) can now be re-written as:

$$z_{hgj} = a_j + \boldsymbol{\theta}_{1j} X_{hg} + b_j(\delta d_U + E_{hg} + u_{hg}) \tag{13}$$

where $a_j = \delta_j + \theta_{2j} \ln C_R$ and $b_j = \theta_{2j}\beta$. I use a generalised structural equation model (GSEM) which implements an IRT model with this form of z_{hgj}. I use 7-point Adaptive Gaussian Quadrature as the method of integration.

Importantly, while $b_p = \theta_{2j}\beta$, it is not possible to identify θ_{2j} and β separately. This is because l_{hg} is a latent variable, and so could have any scale. Stata still allows identification of b_p by normalising the factor loading on the latent variable to be 1 in one of the product equations (e.g. $\theta_{2,j=1} = 1$). This is equivalent to setting σ_l to 1 and and allowing the factor loadings (coefficients on the latent variable) to vary freely.

Implementation: Aside from replicating Young's results with the adapted (IRT) methodology, I also adapt (13) to include more groups (decompose the consumption gap further). The general form is then as follows:

$$z_{hgj} = a_j + \theta_{1j} X_{hg} + b_j((\delta_2 d_2 + \delta_3 d_3 .... + \delta_k d_k) + E_{hg} + u_{hg}) \tag{14}$$

There are k regions instead of only the urban-rural distinction, and each d_k is a dummy variable. Region 1 is the base category included in a_j. Each δ_k represents the consumption gap, net of education, between k and the base category. I use the consumption variables described in section 2 as items, and (following Young (2013)) use the number of members in the household (and the number of children) as the demographic variables. The relevant education (years of education for adults), demographic and consumption items are all included in the DHS which is described in more detail in section 2.

Strengths & Limitations: My approach implements Young as an adaptation of a much simpler model: IRT. The main limitation of my approach is the assumption that $q_{hj}|l,\boldsymbol{\theta},\boldsymbol{X}$ are independently and identically distributed. This is also a limitation in Young's original approach. This assumption allows the co-movement in products to be attributed to changes in log real consumption only.

In terms of independence across households, variation in local prices may cause correlated preference shocks across households. Therefore, the logistically

distributed errors may not be independent implying that $q_{hj}|l,\boldsymbol{\theta},\boldsymbol{X}$ is also not. This concern could be mitigated by clustering households at the DHS cluster level. When Young (2013) did this, he found results were insensitive to clustering.

In terms of independence across products, this essentially means that the model assumes away substitution effects. Young (2012) argues that attempting to find a diverse, large and ultimately representative product sample partly alleviates this concern. He then verifies that the omission of certain products/product categories does not substantially alter his measurements, and thus illustrates the assumption is benign in considering the urban-rural divide. I will use the same check for robustness when adapting his model. Additionally, since transport items (bicycle/motorcycle/car) are likely substitutes, I only include one item, which is 'vehicle with engine'. The ramifications of this choice are discussed shortly.

A final limitation in my and Young's approach is the implicit assumption of monotonicity: each item is assumed to contribute to the co-movement of the latent variable in only one direction. This may be a problem for items like bicycles, which may be owned by wealthier families in rural areas but poorer families in urban areas (Poirier et al., 2020). This concern can be partially addressed by including only items that are wanted regardless of location. Plausibly, most real consumption items in the DHS meet this criterion. For example, having good housing conditions and healthy children seems to meet this criterion uncontroversially. The exception is bicycles, which I consequently exclude and use 'vehicle with engine' as the transport variable instead. This has the disadvantage of perhaps dampening the inequality within rural areas this model can identify.

Principal Component Analysis: The main competitor of Item Response Theory for creating a consumption index is principal component analysis (PCA). Both methods are often used to create wealth indices, which are similar to consumption indices but utilise a smaller item sample. For example, they include only ownership of assets, rather than an inclusive sample composing the majority of budget-share of lower-income households.

PCA was first pioneered as a wealth index by Filmer and Pritchett (2001), who

showed that in India it was effective at predicting education outcomes. It is now commonly adopted (Poirier et al., 2020), with a prominent example being the wealth index created by the DHS itself (Rutsein and Johnson, 2004). The results of IRT and PCA have been found to be highly correlated (Filmer and Scott, 2012).

Conventionally, PCA reduces the dimensionality of consumption variables into a single consumption index. The consumption index is a linear combination of the consumption variables, where the weights on consumption variables are chosen to minimise information lost in the form of variation in the data.

Compared to PCA, my IRT approach has the advantage of having an economic interpretation (although at the cost of some restrictive assumptions). This interpretation is obtained by, like Young (2013), converting inequality into 'education equivalent units' (described previously). Additionally, my approach is easily implemented with clustering and survey weights.

Nevertheless, I implement PCA as a robustness check for these results. I first create a consumption index using PCA and then run a cross-sectional regression of predicted household consumption on region, education and demographics. For simplicity, the second part of the process is weighted by the DHS survey weights and clustered at the location cluster level, but the first part of the process is not.

6 Results, Robustness & Discussion

This section is organised as follows. The first subsection illustrates that I am able to implement Young (2013), the seminal paper on measuring urban-rural consumption gaps, as a modification of a much simpler model (IRT). The adapted IRT model is more transparent and simpler to implement than Young's original approach.

The second subsection extends that methodology to measuring consumption gaps based on decomposed rural and urban areas (categorised by distance to and size of city). I verify, at scale, that living standards increase by proximity to and size of city. Additionally, the smallest urban-rural divide, between the worst-off urban area and the best-off rural one, is approximately two-thirds of the aggregate gap. This goes some way to reconciling the large consumption gaps found in cross-sectional studies with underwhelming panel results – providing an alternative explanation to sorting. I further find that, while the urban-rural divide is the largest regional disparity, within-urban inequality can also be substantial.

The final subsection incorporates informal settlements. Due to the creation of the first transcontinental informal settlement map, living standards in slums can finally be measured and compared to other regions. Identified slums have robustly lower living standards than other areas in the largest city, with the exception of Kenya. However, compared to even the most well-off rural quartile, informal settlements have considerably higher living standards across almost every metric and in every country. The occasional exception is children's health outcomes (potentially because of the higher density of people in cities), however, this does not tend to translate into higher death rates. Consequently, I concur with Gollin et al. (2021): It appears a spatial equilibrium between urban and rural areas in sub-Saharan Africa is unlikely.

6.1 Modifying Young's Approach

I implement Young (2013) as a modification of an IRT model, and report the results compared to Young's original paper in Table 3. My implementation uses the survey weights released with the DHS, and clusters standard errors at the location cluster

level. Despite the analysis being on a smaller sample and in a later time period, the mean results are similar across studies. As in Young's paper, the urban-rural gap accounts for over double the inequality of the regional education gap. In terms of differences in standard deviations, having a more homogeneous sample than Young in terms of region and time period could decrease standard deviation. On the other hand, having a smaller sample could increase it. It seems for $\hat{\delta}$ it decreases, while for the education gap it increases slightly. Table 3 ultimately illustrates that an IRT method is a viable method for implementing Young, while being considerably easier to implement.

Table 3: Young (2013) Results Compared to Results with Approach Implemented as Modification of IRT

	Young (2013) Estimates	**IRT Implementation**
Sample	65 Developing Countries	10 Sub-Saharan Countries
Timing	(1990 – 2013)	(2015 – 2020)
Mean $\hat{\delta}_{urban}$	9.38	9.12
SD of $\hat{\delta}_{urban}$	(4.33)	(3.12)
Mean $E_U - E_R$	3.00	3.43
SD of $E_U - E_R$	(0.91)	(1.10)

Notes: The first column shows the results reported in Young (2013). The second column shows the results estimated via the methodology described in section 5. $\hat{\delta}_{urban}$ is the estimated urban-rural consumption gap (net of education) in education equivalent units. $E_U - E_R$ is the average education in urban areas compared to rural ones. Mean and standard deviations are taken over countries in the sample.

6.2 City-Based Approach to Decomposing Areas

I am then able to decompose the urban-rural gap using a city-based approach. I implement the same (adapted Young/IRT) methodology, but with rural areas disaggregated by distance to cities and urban areas disaggregated by size of city. The results are below in Figure 10 and the left panel of Table 4. In terms of magnitude, on aggregate (Table 4) and for the majority of countries (Figure 10) living standards decrease by decreasing size of city[35] and distance away from city.

[35] 'Other urban areas' would be considered the smallest 'city'.

Although the standard errors are large in Table 4, particularly in urban areas,[36] the consistency of this pattern across countries – as illustrated in Figure 10 – lends more plausibility to the results. This outcome is in line with the urban spillover literature, but to my knowledge has never been verified in sub-Saharan Africa at this scale.

South Africa and Zimbabwe are exceptions to living standards decreasing when going from the largest cities to other cities (Figure 10). In South Africa, this could be because there are multiple similarly sized large cities. By contrast, in most other countries in the sample, there is a large margin between one city, usually the capital, and others. For Zimbabwe, the reason it is an exception is less clear. Regardless, due to the large confidence intervals, not too much should be inferred from individual cases rather than the general trend.

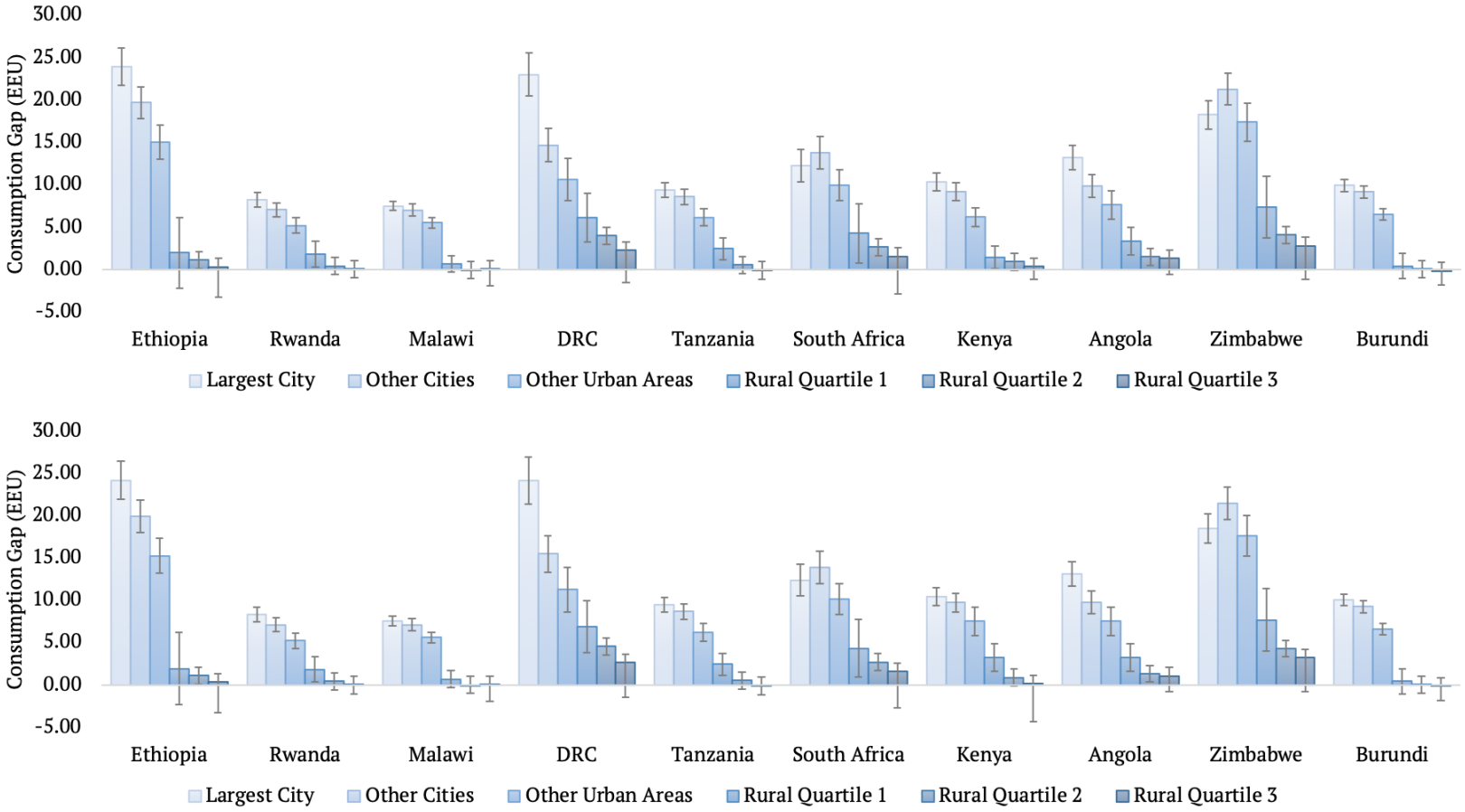
Table 4: Results of City-Based Approach to Disaggregating Urban-Rural Gap

	Includes Health/Education	**Excludes Health/Education**
Mean $\hat{\delta}_{LargestCity}$	13.60	13.81
SD of $\hat{\delta}_{LargestCity}$	(5.70)	(5.92)
Mean $\hat{\delta}_{OtherCities}$	12.02	12.22
SD of $\hat{\delta}_{OtherCities}$	(4.87)	(4.94)
Mean $\hat{\delta}_{OtherUrbanAreas}$	9.01	9.29
SD of $\hat{\delta}_{OtherUrbanAreas}$	(4.00)	(4.03)
Mean $\hat{\delta}_{RuralQuartile1}$	2.99	3.27
SD of $\hat{\delta}_{RuralQuartile1}$	(2.20)	(2.30)
Mean $\hat{\delta}_{RuralQuartile2}$	1.54	1.58
SD of $\hat{\delta}_{RuralQuartile2}$	(1.45)	(1.59)
Mean $\hat{\delta}_{RuralQuartile3}$	0.85	0.89
SD of $\hat{\delta}_{RuralQuartile3}$	(1.01)	(1.14)

Notes: The first column shows the results when the consumption index includes health/education variables. The second column shows the results when the consumption index excludes health/education variables. All results are estimated via the methodology described in section 5. $\hat{\delta}_{Region}$ *is the estimated consumption gap (net of education) between that region and the quartile of rural areas furthest away from the city (Quartile 4). Mean and standard deviations are taken over countries in the sample.*

[36] The higher standard deviation within urban areas compared to rural ones may be due to some of the types of consumption items included in the index being uncommon in rural areas, but available to varying degrees in urban ones.

Figure 10: Country Break-Down of Results of City-Based Approach to Disaggregating Urban-Rural Gap



Notes: The country break-down of results in Table 4. The top panel is when health/education variables are included, the bottom panel is when health/education variables excluded. All consumption gaps are in educational equivalent units (EEU) and relative to the base category of rural areas furthest from the city (Quartile 4). 95% confidence intervals shown. The figures illustrate that living standards tend to increase by proximity to and size of city.

It is also striking, in Figure 10, that there is generally less inequality within urban and rural areas than between them – there is a particularly steep drop in many countries at the urban-rural divide. On aggregate (Table 4), living standards in 'the worst'[37] urban area are ∼6.02 education equivalent units (EEU) higher than the 'the best' rural area.[38] By contrast, maximum within-urban and within-rural inequality is ∼4.59 and ∼2.99 EEU respectively.[39] While excluding bicycles from the sample may dampen the within-rural inequality identified, it is unlikely to halve it. The focus in development economics on the urban-rural gap is vindicated by these results: The largest jump in inequality appears to be between urban and rural areas. Decomposing the gap is able to provide a more nuanced understanding, however, than the previous black boxes of the binary categories 'urban' and 'rural'.

For example, other regional inequality is substantial. The maximum within-urban gap is still higher than regional education inequality as measured in Table 3. Although, the standard errors are large and Figure 10 illustrates substantial heterogeneity between countries (in the extent of urban inequality), such that some countries, specifically Ethiopia and the DRC, may drive this result.

These results also illustrate that the extent of the living standards differential does depend on where in a rural area to where in an urban area is being compared: On aggregate, the maximum urban-rural inequality (∼13.6 EEU) is over double the minimum urban-rural inequality (∼6.02 EEU).[40] For proponents of the 'frictions' paradigm (specifically transport frictions), these results may be particularly important for the cost-benefit analysis of policies. For example, the results imply (in this paradigm) that those who are most able to gain from movement to cities are also those for whom alleviating transport frictions is likely more costly (as they live the furthest from cities).

Additionally, the minimum urban-rural inequality (∼6.02 EEU) is about two-thirds of the inequality of the urban-rural gap on aggregate (∼9.12 EEU, as measured in Table 3). This substantially lower number could go part way to

[37] In terms of estimated living standards.

[38] $\hat{\delta}_{OtherUrbanAreas} - \hat{\delta}_{RuralQuartile1}$ (left panel of Table 4)

[39] $(\hat{\delta}_{LargestCity} - \hat{\delta}_{OtherUrbanAreas})$ and $(\hat{\delta}_{RuralQuartile1} - \hat{\delta}_{RuralQuartile4})$ respectively.

[40] $(\hat{\delta}_{LargestCity} - \hat{\delta}_{RuralQuartile4})$ and $(\hat{\delta}_{OtherUrbanAreas} - \hat{\delta}_{RuralQuartile1})$ respectively

reconciling the underwhelming panel data results with the cross-sectional measurement of gaps, and provide an alternative explanation to sorting.

For robustness, following Young (2013), I check that the assumption of no substitution effect is innocuous by verifying that the results are robust to omitting product categories. Specifically, I repeat the exercise while leaving out the health/education category. I choose this category since it seems plausible that people are substituting assets and/or housing conditions for better access to services. The results are robust to this. Both Table 4 and Figure 10 show that this product category does not substantially change results or standard errors, for any country. Finally, I measure inequality by using a simple PCA model, and find that this illustrates the same pattern of higher living standards by proximity to and size of city (see Table A1 in appendix).

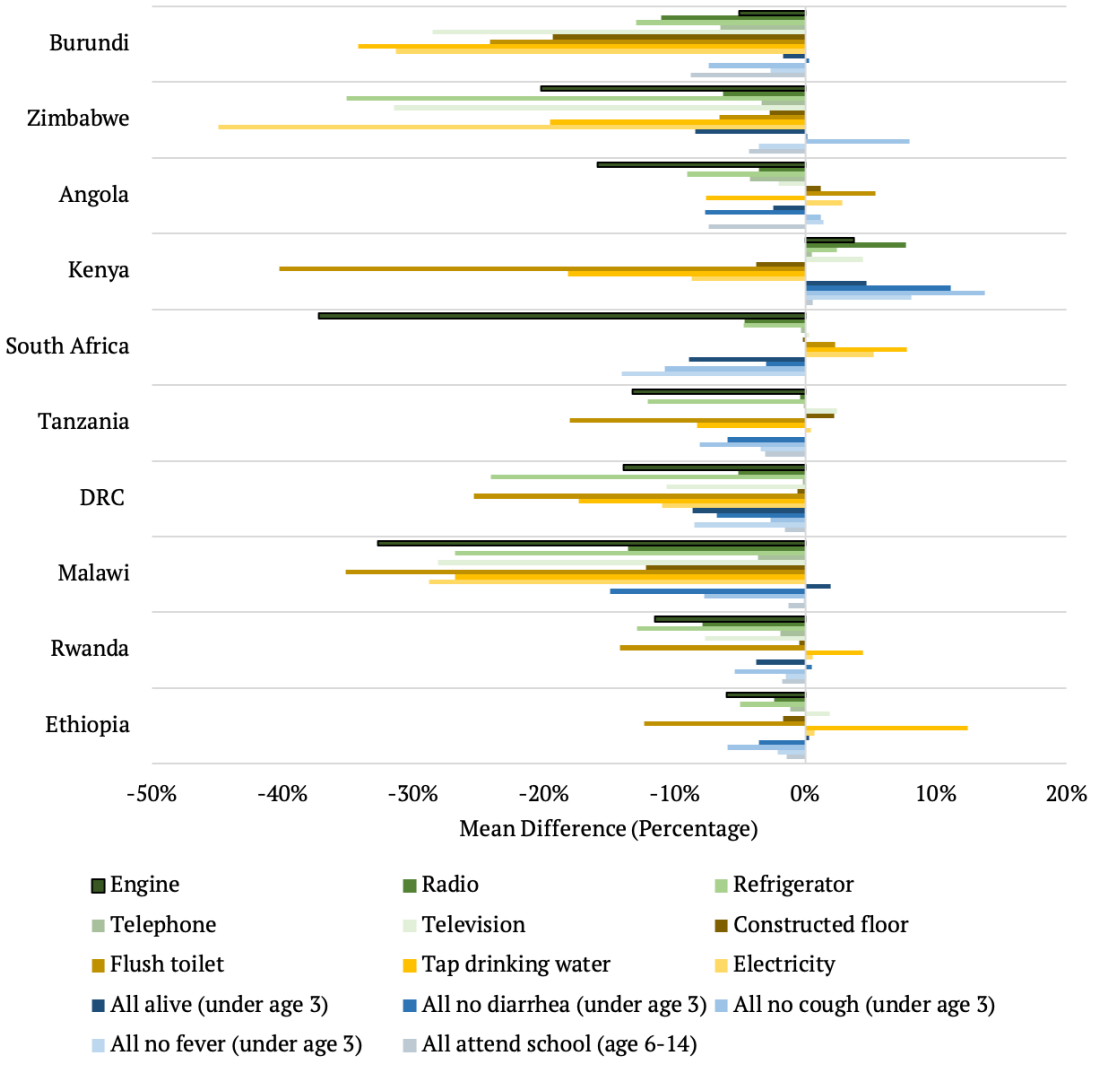
6.3 Incorporating Informal Settlements

The previous results imply that living standards in the largest city are the best in the country for the vast majority of the sample. This subsection evaluates whether there are some areas in the largest city where this is not the case, specifically, informal settlements.

Informal Settlements Compared to Rest of City: I verify that informal settlements identified in this paper tend to have lower living standards than the rest of the largest city. Table 5 (left panel) illustrates that for every standard of living measure slums are on average worse-off by 3%-15%.

Figure 11 shows the country breakdown. To my knowledge, this is the first time living standards across slums can be compared at this scale. The pattern across countries is generally consistent: In almost every country slums are 'worse-off' on most consumption measures. The exception is Kenya, where while all housing condition measures are lower in slums, health/education and asset ownership indicators are higher. This is consistent with substitution in Kenya, with people living in informal housing to access better amenities/opportunities. This could be the 'pull' factor that attracts people into slums.

Figure 11: Difference in Mean (of Consumption Indicators) Within Largest City $(\hat{\mu}_{Slum} - \hat{\mu}_{ExcludingSlum})$



Notes: Country break-down of results in Table 5 (left panel). Each product category is represented by a different colour: Asset Ownership (green), Health/Education (blue) and Housing Conditions (brown/yellow). The largest city breakdown is $\hat{\mu}_{Slum} - \hat{\mu}_{ExcludingSlum}$, where $\hat{\mu}$ is mean of the consumption indicator in a country. The figure illustrates slums identified in this paper tend to have substantially lower living standards than the rest of the city.

Table 5: Difference in Mean (of Consumption Indicators) Comparisons Aggregated Across Countries

	Largest City Breakdown $\hat{\mu}_{Slum} - \hat{\mu}_{ExcludingSlum}$		**Slum Breakdown** $\hat{\mu}_{Irregular} - \hat{\mu}_{Regular}$		**Slum vs Rural Quartile 1** $\hat{\mu}_{Slum} - \hat{\mu}_{RuralQuartile1}$	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Health & Education						
All attend school (age 6-14)	-0.03	0.001	-0.01	0.002	0.09	0.006
All no fever (under age 3)	-0.03	0.003	0.00	0.007	0.06	0.010
All no cough (under age 3)	-0.03	0.006	-0.02	0.014	0.00	0.010
All no diarrhea (under age 3)	-0.03	0.004	-0.03	0.011	0.01	0.010
All alive (under age 3)	-0.03	0.002	-0.01	0.003	0.08	0.003
Housing Conditions						
Electricity	-0.11	0.027	0.02	0.050	0.53	0.055
Tap drinking water	-0.11	0.021	-0.02	0.040	0.33	0.047
Flush toilet	-0.17	0.020	0.03	0.014	0.35	0.058
Constructed floor	-0.04	0.004	0.04	0.008	0.55	0.067
Asset Ownership						
Television	-0.10	0.018	0.01	0.016	0.44	0.053
Telephone	-0.02	0.000	0.02	0.003	0.30	0.038
Refrigerator	-0.14	0.012	0.01	0.021	0.24	0.029
Radio	-0.05	0.003	0.01	0.007	0.17	0.014
Engine	-0.15	0.014	-0.01	0.007	0.05	0.001

Notes: $\hat{\mu}$ is mean of consumption indicator in a specific country. Slum breakdown is between two equally sized groups based on regular/irregular layout. Mean across countries rounded to two decimal places, standard deviation to three decimal places. Consumption indicators as described in detail in section 2.

It is also worth noting that in South Africa there is potentially the inverse pattern (although less clear): some housing conditions are better, but health/education and asset ownership is worse.

Since many of these consumption indicators may be sensitive to the number of members and children in the household, I repeat the above analysis but with cross-sectional regressions on demographics (see Table A2 and Figure A3 in the appendix). The patterns described in this subsection are robust to this, with the exception that South Africa's outcomes are even less clearly the inverse of Kenya's pattern.

Within-Slum Inequality: To proxy for planning/formality, I measure the difference in living standards between irregularly and regularly laid out slums.[41] Results are displayed in Figure 12 and the middle panel of Table 5. There is no apparent consistent pattern in these results: Regular layout on aggregate (Table 5) is positively associated with some consumption measures and negatively with others. Overall, the magnitude of these differences is not large.

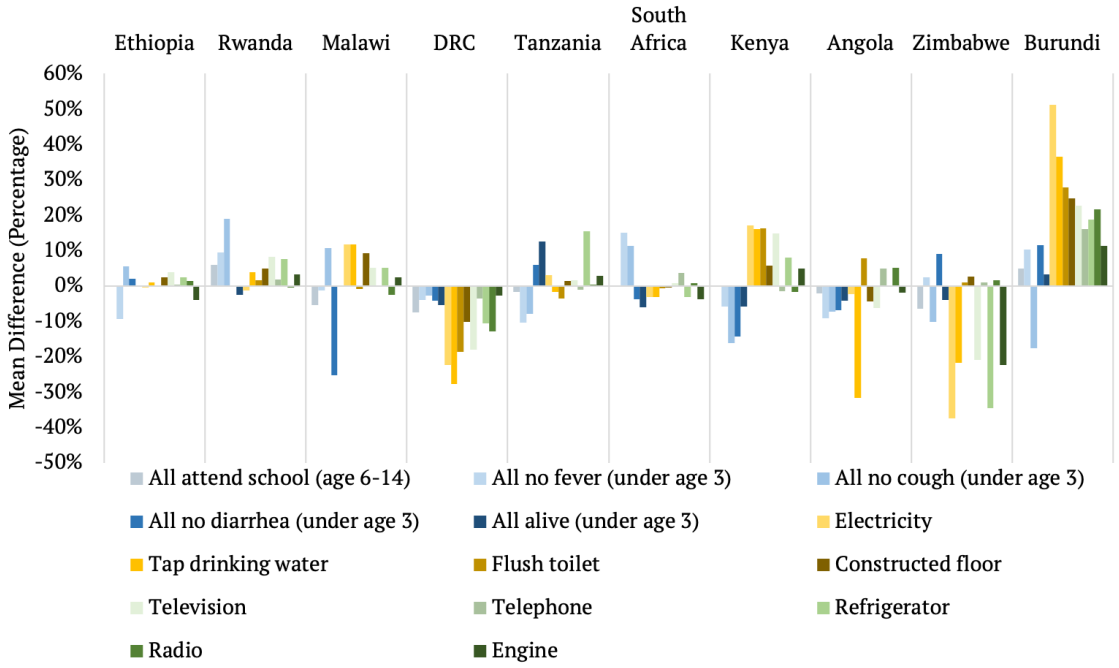
The country breakdown (Figure 12) shows substantial heterogeneity. In a minority of countries, a vast majority of consumption indicators are strongly correlated with irregular layout. In Zimbabwe and the DRC, this correlation is negative (many consumption indicators are lower for more informal housing). In Burundi, it is positive.

These results persist when (for robustness) I use a cross-sectional regression on demographic variables (see Table A2 and Figure A3 in the appendix).

Theoretically, these patterns could relate to whether gravitating toward very informal housing in the city is a consequence of 'push' (like civil wars/natural disasters) or 'pull' factors. The former could create a negative correlation between irregular layout (informality) and consumption, and the latter could create a positive one. However, this sample may be too small to test this empirically. Consequently, I leave this to further research.

[41] With the two groups equally sized and separated by the median.

Figure 12: Difference in Mean (of Consumption Indicators) Within Slums $(\hat{\mu}_{Irregular} - \hat{\mu}_{Regular})$



Notes: The country break-down of results in middle panel of Table 5. Each product category is represented by a different colour: Asset Ownership (green), Health/Education (blue) and Housing Conditions (brown/yellow). The slum breakdown is by how irregularly laid out buildings are: $\hat{\mu}_{Irregular} - \hat{\mu}_{Regular}$, where $\hat{\mu}$ is mean of the consumption indicator in a specific country.

Informal Settlements and the Urban-Rural Gap: Finally, I attempt to incorporate slums into the adapted IRT/Young approach for measuring inequality. Results are below in Table 6. Table 6 is the counterpart of Table 4, but now with the largest city decomposed into informal settlements and the rest of the city. Compared to Table 4, inclusion of informal settlements does not substantially change the values of other regions. As expected, the living standards of the largest city, now excluding informal settlements, increases in magnitude.

When health/education is included, the IRT/Young method reflects slightly lower living standards – closer to the living standards of smaller cities – for those living in slums compared to the rest of the largest city. However, when health/education is excluded, the resulting asset index reflects living standards in informal settlements considerably lower than the largest city – and often

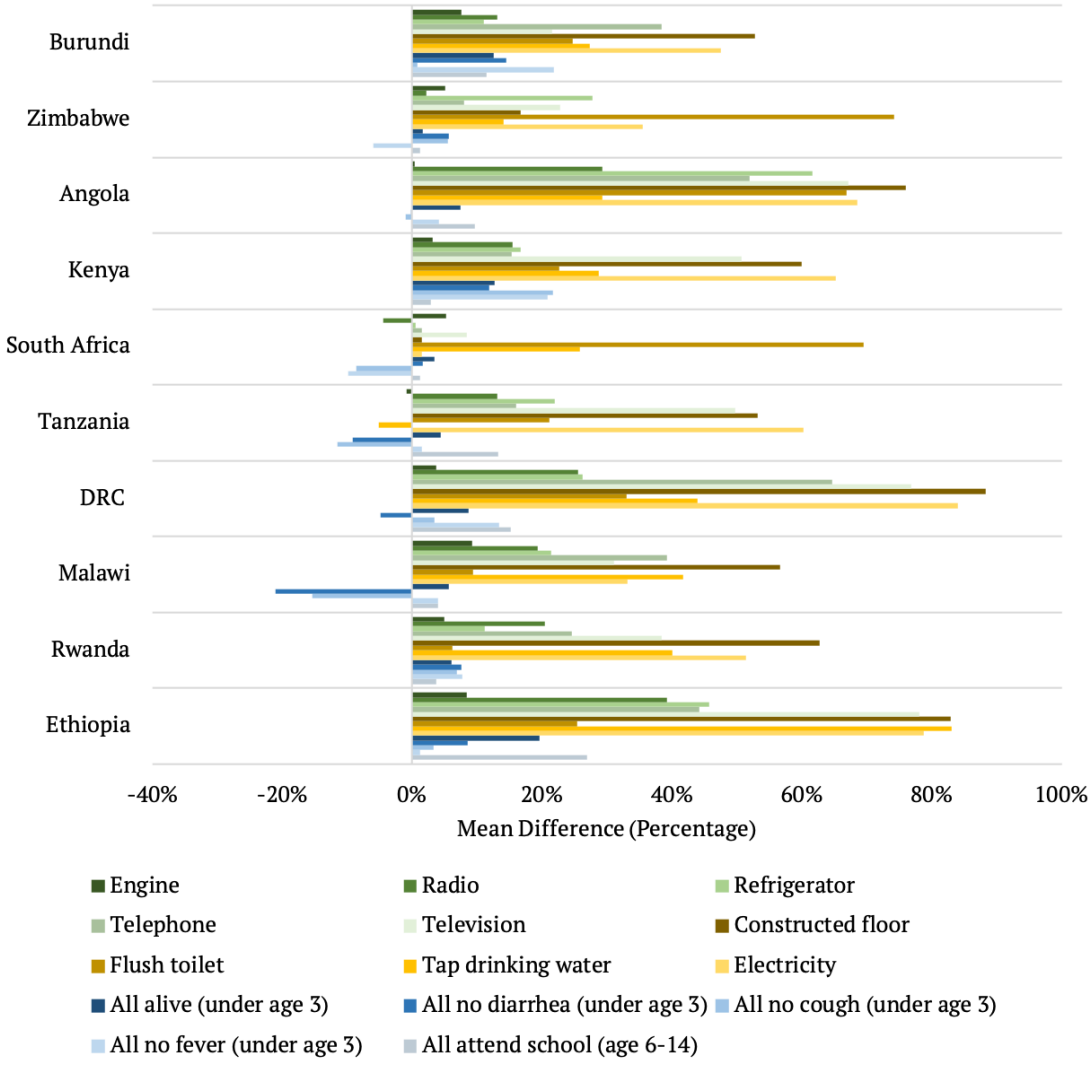
overlapping with rural areas. This result is reflected both on aggregate (Table 6) and for the majority of countries in the sample (Figure A2 in appendix). Consequently, the assumption underlying the adapted IRT method (of no substitution effects) is inappropriate in this context. This is a fundamental assumption of both the IRT approach and (implicitly) PCA, and consequently these approaches are abandoned. Since it is no longer possible to create appropriate consumption indices, I consider individual consumption indicators instead.

Table 6: Consumption Index (IRT Approach) Results on Inclusion of Informal Settlements

	Includes Health/Education	**Excludes Health/Education**
Mean $\hat{\delta}_{InformalSettlement}$	12.70	5.22
SD of $\hat{\delta}_{InformalSettlement}$	(6.04)	(5.33)
Mean $\hat{\delta}_{OtherLargestCity}$	14.47	14.69
SD of $\hat{\delta}_{OtherLargestCity}$	(5.86)	(6.08)
Mean $\hat{\delta}_{OtherCities}$	12.13	12.39
SD of $\hat{\delta}_{OtherCities}$	(4.97)	(5.04)
Mean $\hat{\delta}_{OtherUrbanAreas}$	9.09	9.23
SD of $\hat{\delta}_{OtherUrbanAreas}$	(4.10)	(4.16)
Mean $\hat{\delta}_{RuralQuartile1}$	3.00	3.09
SD of $\hat{\delta}_{RuralQuartile1}$	(2.24)	(2.37)
Mean $\hat{\delta}_{RuralQuartile2}$	1.54	1.61
SD of $\hat{\delta}_{RuralQuartile2}$	(1.48)	(1.59)
Mean $\hat{\delta}_{RuralQuartile3}$	0.85	0.88
SD of $\hat{\delta}_{RuralQuartile3}$	(1.03)	(1.10)

Notes: The first column shows the results when consumption index includes health/education variables. The second column shows the results when the consumption index excludes health/education variables. All results are estimated via the methodology described in section 5. $\hat{\delta}_{Region}$ *is the estimated consumption gap (net of education) between that Region and the quartile of rural areas furthest away from the city (Quartile 4). Mean and standard deviations are taken over countries in the sample.*

Figure 13: Difference in Mean (of Consumption Indicators) Between Slums and Most Well-off Rural Quartile ($\hat{\mu}_{Slum} - \hat{\mu}_{RuralQuartile1}$)



Notes: The country break-down of results in right panel of Table 5. Each product category is represented by a different colour: Asset Ownership (green), Health/Education (blue) and Housing Conditions (brown/yellow). The comparison is $\hat{\mu}_{Slum} - \hat{\mu}_{RuralQuartile1}$, where $\hat{\mu}$ is mean of the consumption indicator in a specific country. This figure illustrates that slums tend to be better off than even the most well-off (closest to the city) rural area.

I compare the 'best off' rural area (closest to the city) to areas in informal settlements. Table 5 (right panel) and Figure 13 display these results. Even when compared to the best rural areas, on aggregate (Table 5) informal settlements are at least as well-off, and often considerably better-off, on every consumption index. The country breakdown in Figure 13 largely supports these results. However, it shows that in certain countries some outcomes are slightly worse in slums. The most common/consistent outcome to be worse is some health outcomes (disease) for children. This result persists even when (for robustness) I use the cross-sectional regression on demographics (see Figure A3 and Table A2 in appendix).

This would imply that, in over half the countries, certain consumption indicators are higher and certain lower in informal settlements (compared to rural areas). Consequently, the result of Gollin et al. (2021) that all living standards increase monotonically with population density seems to break down for slums in these countries. The distribution of people between the (best-off) rural area and the slum could be a spatial equilibrium, so long as the individual weighted certain children's health outcomes sufficiently highly.

However, it should be noted that these worse health outcomes in terms of disease do not seem to translate into higher (children's) deaths in slums compared to rural areas (except in Tanzania, in the cross-sectional regression – see Figure A3). It could be that living standards in informal settlements, including the density of people, facilitate the easier spread of disease. However, a combination of better services (or being better off in other regards) equips households to combat the higher disease burden. Therefore, although the monotonicity in Gollin et al. (2021) does not technically hold in this case, it remains unlikely that there is a spatial equilibrium. Rather, living standards in the largest city appear to be better than even the best rural area, even when considering living standards in slums.

7 Conclusion

Despite the urban-rural consumption gap being a substantial component of inequality in the developing world, it has remained largely a black box. This paper explores regional inequality in finer detail than was previously possible in sub-Saharan Africa.

To measure living standards, I improve upon the implementation of the seminal paper measuring the rural-urban consumption gap: Young (2013). I illustrate that Young's approach is similar to a much simpler model, and can be implemented as an adaption of Item Response Theory. The adapted approach has numerous advantages over the original, including: (i) it is simpler to implement since many programs (including Stata) implement IRT models with a single line of code; and (ii) it increases transparency, which was previously considered a downfall Young's original method (Lagakos, 2020). Further, the popularity of PCA for constructing consumption indices, despite its lack of economic interpretation, may partially be attributed to its straightforward implementation. The developments in this paper could allow the adapted Young approach to compete against other indices.

To connect living standards with spatial data in sub-Saharan Africa, I develop improvements to the current best practice for connecting displaced survey data to neighbourhoods. Consequently, it was possible to identify households more likely to live in informal settlements – a task previously considered infeasible (Gollin et al., 2021). In addition, the probability weighting technique I develop could be used in future research to map amenities across a city. This would allow the mapping of living standards and/or provide a proxy for income. Finally, these measures could be used to effectively allocate resources in the policy realm.

Additionally, to include informal settlements in the decomposition of living standards, I created the first transcontinental map of slums. Despite informal settlements becoming an increasingly representative aspect of urban life in sub-Saharan Africa, research at scale on outcomes/dynamics/living conditions within them has been held back by not having a map of slums based on a geographically consistent definition.

To proxy for planning within slums, I developed a tool to measure how regularly laid out buildings are. This tool measures the deviation of building walls from orthogonal axes, where the axes are chosen to minimise that deviation. I then scaled this measure over half a million $100m^2$ areas. In future research, this technique could be used to proxy for planning in other settings (for example, city planning).

These innovations allowed the consumption gap to be decomposed using a city-based approach, and including informal settlements. I find that living standards increase monotonically by distance to and size of city. The urban-rural gap is the largest source of regional inequality, but within-urban inequality is still substantial – on average a larger component of total inequality than the urban-rural education gap. Additionally, I find that the minimum urban-rural gap is two-thirds of the urban-rural gap on aggregate.

With regard to slums, informal settlements have robustly lower living standards than their counterparts in the largest city, with the exception of Kenya. However, compared to even the most well-off rural quartile, I find that informal settlements have considerably higher living standards across almost every metric and in every country. The occasional exception is children's health outcomes (potentially because of the higher density of people in cities). However, this does not tend to translate into higher death rates. Consequently, despite slums potentially being poverty traps (Marx et al., 2013), they are unlikely to be a bottleneck for migration.

These results show sub-Saharan urbanisation bucks the historical trend: Urban areas no longer only reflect higher wages, but also higher overall living standards, even in slums.

References

J. A. Alvarez. The Agricultural Wage Gap: Evidence from Brazilian Micro-Data. *American Economic Journal: Macroeconomics*, 12(1):153–73, 2020.

F. Caselli. Accounting for Cross-Country Income Differences. *Handbook of Economic Growth*, 1:679–741, 2005.

D. L. Costa and M. E. Kahn. Public Health and Mortality: What Can We Learn from the Past? *Public Policy and the Income Distribution*, pages 359–98, 2006.

D. Filmer and L. H. Pritchett. Estimating Wealth Effects Without Expenditure Data or Tears: an Application to Educational Enrollments in States of India. *Demography*, 38(1):115–132, 2001.

D. Filmer and K. Scott. Assessing Asset Indices. *Demography*, 49(1):359–392, 2012.

D. Gollin, D. Lagakos, and M. E. Waugh. The Agricultural Productivity Gap. *The Quarterly Journal of Economics*, 129(2):939–993, 2014.

D. Gollin, M. Kirchberger, and D. Lagakos. Do Urban Wage Premia Reflect Lower Amenities? Evidence from Africa. *Journal of Urban Economics*, 121:103301, 2021.

M. Greenstone, R. Hornbeck, and E. Moretti. Identifying Agglomeration Spillovers: Evidence from Winners and Losers of Large Plant Openings. *Journal of Political Economy*, 118(3):536–598, 2010.

J. H. Hicks, M. Kleemans, N. Y. Li, and E. Miguel. Reevaluating agricultural productivity gaps with longitudinal microdata. *Journal of the European Economic Association*, 19(3):1522–1555, 2021.

M. S. Johnson. Marginal Maximum Likelihood Estimation of Item Response Models in R. *Journal of Statistical Software*, 20:1–24, 2007.

L. Kesztenbaum and J.-L. Rosenthal. The Democratization of Longevity: How the Poor Became Old in Paris, 1880–1913. In *New Approaches to Death in Cities during the Health Transition*, pages 137–154. Springer, 2016.

P. Kline and E. Moretti. Local Economic Development, Agglomeration Economies, and the Big Push: 100 Years of Evidence from the Tennessee Valley Authority. *The Quarterly journal of economics*, 129(1):275–331, 2014.

A. Kraay and D. McKenzie. Do Poverty Traps Exist? Assessing the Evidence. *Journal of Economic Perspectives*, 28(3):127–48, 2014.

D. Lagakos. Urban-Rural Gaps in the Developing World: Does Internal Migration Offer Opportunities? *Journal of Economic perspectives*, 34(3):174–92, 2020.

A. Marshall. *Principles of Economics: Unabridged Eighth Edition*. Cosimo, Inc., 1890.

B. Marx, T. Stoker, and T. Suri. The Economics of Slums in the Developing World. *Journal of Economic perspectives*, 27(4):187–210, 2013.

M. McMillan, D. Rodrik, and Í. Verduzco-Gallo. Globalization, structural change, and productivity growth, with an update on africa. In *World Development*, volume 63, pages 11–32, 2014.

G. Michaels, D. Nigmatulina, F. Rauch, T. Regan, N. Baruah, and A. Dahlstrand. Planning Ahead for Better Neighborhoods: Long-run Evidence from Tanzania. *Journal of Political Economy*, 129(7):2112–2156, 2021.

C. Perez-Haydrich, J. L. W. Warren, C. Burgert, and M. Emch. Guidelines on the Use of DHS GPS Data. *DHS Spatial Analysis Reports*, 8, 2013.

M. J. Poirier, K. A. Grépin, and M. Grignon. Approaches and Alternatives to the Wealth Index to Measure Socioeconomic Status Using Survey Data: a Critical Interpretive Synthesis. *Social Indicators Research*, 148(1):1–46, 2020.

D. Restuccia, D. T. Yang, and X. Zhu. Agriculture and Aggregate Productivity: A Quantitative Cross-Country Analysis. *Journal of monetary economics*, 55(2): 234–250, 2008.

J. C. Riley. *Rising Life Expectancy: a Global History*. Cambridge University Press, 2001.

S. Rutsein and K. Johnson. The DHS Wealth Index. *DHS Comparative Reports*, 1 (6):115–132, 2004.

W. Sirko, S. Kashubin, M. Ritter, A. Annkah, Y. S. E. Bouchareb, Y. Dauphin, D. Keysers, M. Neumann, M. Cisse, and J. Quinn. Continental-Scale Building Detection from High Resolution Satellite Imagery. *arXiv preprint arXiv:2107.12283*, 2021.

UN-Habitat. *State of the World's Cities 2012/2013: Prosperity of Cities*. Routledge, 2013.

D. Vollrath. The Efficiency of Human Capital Allocations in Developing Countries. *Journal of Development Economics*, 108:106–118, 2014.

J. L. Warren, C. Perez-Heydrich, C. R. Burgert, and M. E. Emch. Influence of Demographic and Health Survey Point Displacements on Point-in-Polygon Analyses. *Spatial demography*, 4(2):117–133, 2016.

J. G. Williamson. Was the Industrial Revolution Worth It? Disamenities and Death in 19th Century British Towns. *Explorations in Economic History*, 19(3):221–245, 1982.

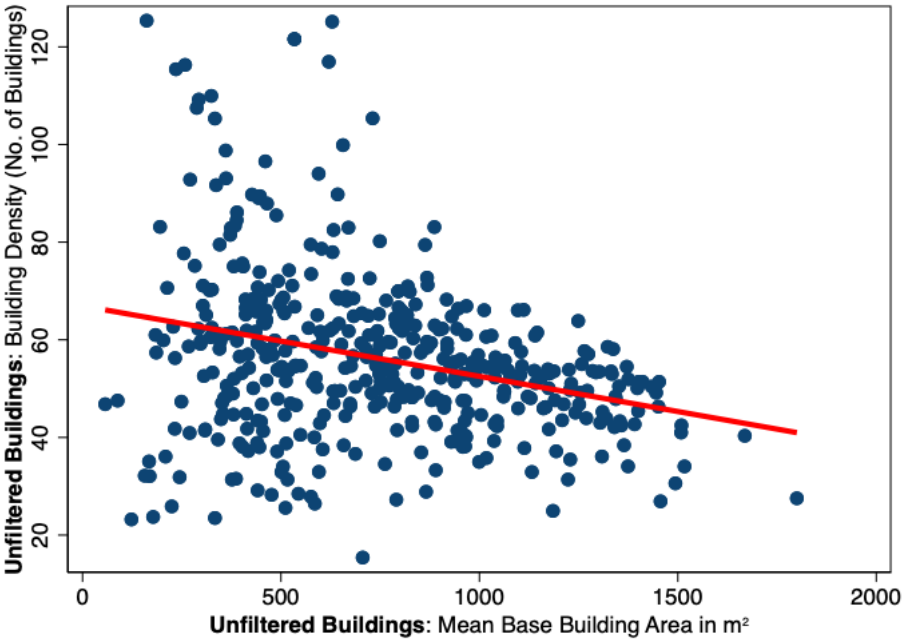
World Bank. Population. Available at `https://data.worldbank.org/indicator/SP.POP.TOTL` (2020).

A. Young. The African Growth Miracle. *Journal of Political Economy*, 120(4): 696–739, 2012.

A. Young. Inequality, the Urban-Rural Gap, and Migration. *The Quarterly Journal of Economics*, 128(4):1727–1785, 2013.

A Appendix

Figure A1: Correlation Between Mean and Density in Largest Cities Across Sample



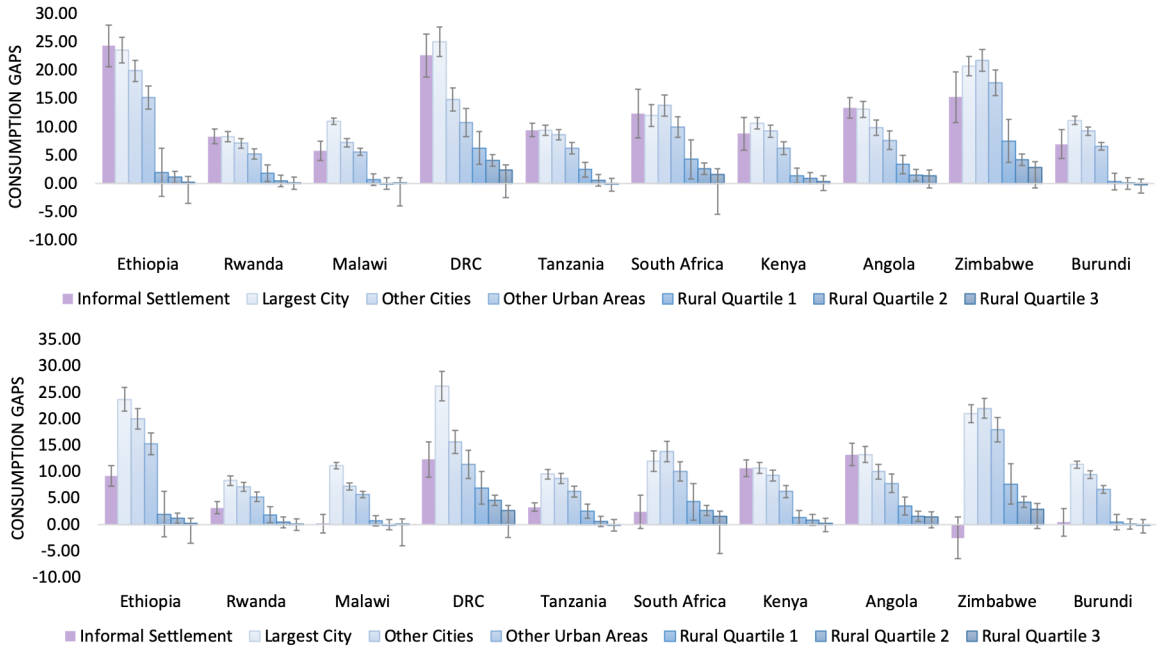
Notes: The Figure shows the correlation (-0.29) between mean and density in the largest cities across the sample. Both measures are unfiltered, and probability weighting (combining the maps in Figure 6 with the Probability Mass Template) is used to get mean and density.

Table A1: PCA: Results of City-Based Approach to Disaggregating Urban-Rural Gap

	PCA Estimate
Mean $\hat{\delta}_{LargestCity}$	3.04
SD of $\hat{\delta}_{LargestCity}$	(0.96)
Mean $\hat{\delta}_{OtherCities}$	2.46
SD of $\hat{\delta}_{OtherCities}$	(0.73)
Mean $\hat{\delta}_{OtherUrbanAreas}$	1.58
SD of $\hat{\delta}_{OtherUrbanAreas}$	(0.57)
Mean $\hat{\delta}_{RuralQuartile1}$	0.34
SD of $\hat{\delta}_{RuralQuartile1}$	(0.33)
Mean $\hat{\delta}_{RuralQuartile2}$	0.07
SD of $\hat{\delta}_{RuralQuartile2}$	(0.18)
Mean $\hat{\delta}_{RuralQuartile3}$	0.03
SD of $\hat{\delta}_{RuralQuartile3}$	(0.16)

Notes: The results when consumption index created by PCA (and includes health/education variables). The index does not have an interpretation like Young's approach, however, one can see that living standards by this index also increase by proximity and size of cities. Results also limited in that did not use weighting or clustering for creating weights on items. All results are estimated via the methodology described in section 5. $\hat{\delta}_{Region}$ *is the estimated consumption gap (net of education) between that region and the quartile of rural areas furthest away from the city (Quartile 4). Mean and standard deviations are taken over countries in the sample.*

Figure A2: Country Break-Down of Results when Informal Settlements Included



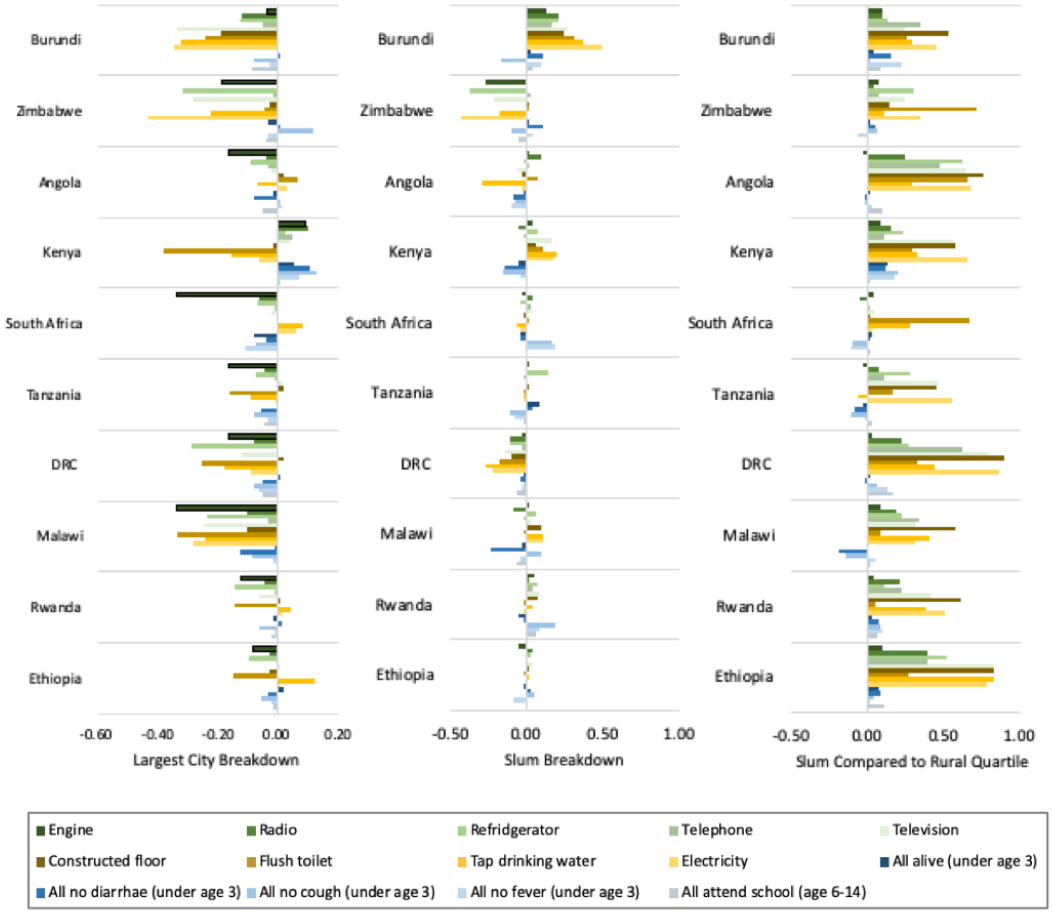
Notes: The country break-down of results in Table 6. Top panel is when health/education variables included, bottom panel is when health/education variables excluded. The category 'Largest City' now excludes informal settlements. All consumption gaps in educational equivalent units and relative to the base category of rural areas furthest from the city (Quartile 4). 95% confidence intervals shown.

Table A2: Difference in Mean (of Consumption Indicators) Comparisons Aggregated Across Countries – with Demographic Controls

	Largest City Breakdown $\hat{\mu}_{Slum}$		**Slum Breakdown** $\hat{\mu}_{Irregular_Layout}$		**Slum vs Rural Quartile 1** $\hat{\mu}_{Slum}$	
	Mean	**SD**	**Mean**	**SD**	**Mean**	**SD**
Health & Education						
All attend school (age 6-14)	-0.03	0.001	-0.01	0.002	0.06	0.003
All no fever (under age 3)	-0.02	0.002	0.01	0.008	0.05	0.010
All no cough (under age 3)	-0.03	0.006	-0.01	0.015	0.01	0.010
All no diarrhea (under age 3)	-0.03	0.003	-0.03	0.011	0.02	0.009
All alive (under age 3)	-0.01	0.001	-0.01	0.002	0.03	0.002
Housing Conditions						
Electricity	-0.11	0.028	0.00	0.053	0.51	0.055
Tap drinking water	-0.10	0.020	-0.01	0.039	0.33	0.048
Flush toilet	-0.16	0.018	0.03	0.014	0.35	0.053
Constructed floor	-0.03	0.004	0.04	0.008	0.54	0.070
Asset Ownership						
Television	-0.10	0.017	0.02	0.018	0.45	0.057
Telephone	-0.01	0.001	0.02	0.003	0.27	0.036
Refrigerator	-0.14	0.010	0.01	0.023	0.27	0.031
Radio	-0.04	0.003	0.02	0.008	0.16	0.014
Engine	-0.15	0.015	-0.01	0.010	0.05	0.002

Notes: $\hat{\mu}_{region}$ *is coefficient on region in cross-sectional regression of consumption indicator on demographics and region. Base categories respectively are: Capital Excluding Slums, Slums with Regular Layout, Rural Area Quartile 1 (closest to city). Slum breakdown is between two equally sized groups based on regular/irregular layout. Mean across countries rounded to two decimal places, standard deviation to three decimal places. Consumption indicators as described in detail in section 2.*

Figure A3: Difference in Mean (of Consumption Indicators) Comparisons – with Demographic Controls



Notes: The country break-down of results in Table A2. Each product category is represented by a different colour: Asset Ownership (green), Health/Education (blue) and Housing Conditions (brown/yellow). Breakdowns correspond to Table A2: The largest city breakdown is $\hat{\mu}_{Slum}$ *(base category rest of city), slum breakdown is* $\hat{\mu}_{Irregular}$ *(base category regularly laid out slums), and third panel is* $\hat{\mu}_{Slum}$ *(base category rural quartile 1).*