Assessing the Accuracy of Proxies of Economic Activity

Prachi Jhamb

Department of Agricultural and Applied Economics, University of Georgia, pj40553@uga.edu

Susana Ferreira

Department of Agricultural and Applied Economics, University of Georgia, sferreir@uga.edu

Patrick Stephens

Department of Integrative Biology, Oklahoma State University, patrick.stephens@okstate.edu

Mekala Sundaram

Department of Integrative Biology, Oklahoma State University, mekala.sundaram@okstate.edu

Jonathan Wilson

Department of Pathology, University of Georgia, jonathan.wilson@uga.edu

(Preliminary Draft. Please Do Not Cite.)

Selected Paper prepared for presentation at the 2023 Agricultural & Applied Economics Association Annual Meeting, Washington DC; July 23-25, 2023

Assessing the Accuracy of Proxies of Economic Activity[1]

Prachi Jhamb; Susana Ferreira, Patrick Stephens, Mekala Sundaram, Jonathan

Abstract

Accurate understanding of socioeconomic realities in developing countries is hindered by the lack of availability and quality of subnational data. Nationally representative surveys are time-consuming and expensive, leading researchers to explore alternative data sources such as night lights, satellite imagery, and mobile phone records. Night lights, in particular, have been used as a proxy for economic activity due to their correlation with the level of economic activity. However, there is a need to assess the accuracy and validity of these alternative data sources, especially at small spatial scales and their ability to capture economic activity.

This paper aims to address these knowledge gaps by comparing spatial variation in economic activity indices across 34 sub-Saharan African countries. The study tests the accuracy of these proxies in predicting economic activity as measured by our indicators of wealth.

The findings indicate that night lights, including the new harmonized dataset, are a good predictor of wealth index within countries at subnational levels, even after accounting for population density. However, for alternative dependent variables like the Human Development Index (HDI), night lights perform better at explaining variation across countries rather than within countries. These insights contribute to a better understanding of the strengths and limitations of using night lights and other alternative data sources for socioeconomic analysis, particularly in developing countries. The research has implications for various disciplines, including economics, health sciences, and disease ecology, by enabling researchers to incorporate spatial variation in poverty and economic activity into their work more effectively.

[1] Prachi Jhamb is a PhD student and Susana Ferreira is a Professor in the Department of Applied Economics at University of Georgia. Patrick Stephens is an Assistant Professor and Mekala Sundaram is a Post-Doc fellow in the Department of Integrative Biology at Oklahoma State University. Jonathan Wilson is a PhD student in the Department of Pathology at University of Georgia.

1. Introduction

National-level indicators of economic and social progress may provide a bird's eye view of a country's performance, but they often mask the heterogeneity that exists within countries. Subnational variations in development outcomes, resource endowments, etc. can have a significant impact on people's well-being, but their analysis is frequently overlooked due to the lack of reliable disaggregated administrative data. This knowledge gap not only limits our understanding of the drivers of development at the local level but also hinders the effectiveness of policy interventions aimed at addressing specific challenges in particular regions. The need for subnational data in developing countries is particularly critical in the context of infectious disease spillovers and climate change. The spatial heterogeneity in socio-economic conditions, disease transmission patterns, and healthcare infrastructure, within a country can have a significant impact on the effectiveness of spillover response measures. Similarly, in the context of climate change, whose impacts can vary across regions within the same country due to factors such as topography, climate variability, or land uses, subnational data can help in identifying areas that are particularly vulnerable to climate impacts and inform resilience strategies. Without subnational data, researchers in the past have relied on national level development indicators which is problematic since development indicators have significant intra-national variation, particularly in large countries (Kummu et al. 2018).

While researchers can rely on subnational administrative data for developed countries, for which there is typically good availability of fine-grained data, the same is not true for developing countries, where the availability and quality of subnational data can be a major hindrance in accurately understanding the social, economic, and political realities on the ground. Nationally representative surveys are time-consuming and expensive and developing countries are therefore

less likely to be covered by such extensive surveys. In fact, many developing countries don't have their estimates updated for decades (Blumenstock 2016). Surveys, such as those from the Demographic Health Surveys (DHS) program have limited repeated observations of the same location (and even the same country) over time (Weidmann & Theunissen, (2021); World Development Report, 2021). Infact, Yeh et al. (2020) estimate that a given African household would appear in a household survey once in 1,000 years, making it difficult to measure changes in well-being over time at a local level. For this reason, researchers have turned to alternative and non-traditional sources of data such as Nightlights (NTL), daytime satellite imagery or mobile phone Call Detail Records (CDR) (Blumenstock et al. 2015; Steele et al. 2017; Steele et al. 2021).

Night lights have been used as a proxy for economic activity in various studies based on the assumption that the amount of light at night is closely tied to the level of economic activity (most consumption and investment activities that occur during the evening or night require lighting which means that the brightness and changes in brightness of night lights can be used as an indicator of economic activity or growth over time). This has led researchers to use the intensity of night lights not only as a measure of intensity of economic activity but also in other applications such as to study inequality, and testing the accuracy of official statistics across different political contexts (Bundervoet et al. 2015).

While daytime imagery is still very rarely used in economics literature, the use of nightlights data, on the other hand, has been increasing ever since its use in an article by Henderson et al. (2011) in the American Economic Review (Gibson et al. 2021, Goldblatt et al. 2020). A search on IDEAS/RePEc in 2023 (using terms such as "nightlights" or "nighttime luminosity" or "nighttime lights (NTL)"), yielded over 230 articles and papers, 134 of them since 2020.

However, as noted by Gibson (2021), a significant problem with the use of night lights in economics research is the widespread use of outdated and inaccurate Defense Meteorological Satellite Program (DMSP) data, the production of which stopped in 2013. DMSP data suffers from various inaccuracies, including blurred images, geo-location errors, and top-coding, leading to the light being attributed to incorrect locations and also making it challenging to distinguish between low-light intensity and high-light intensity areas as the two are grouped together. A review of 41 economics articles and working papers published in 2019 or 2020 that utilized nightlights data, found that over 90 percent of them still used DMSP data (Gibson et al. 2021). This raises concerns about the reliability of published economic results based on this type of data. Superior/ more precise and newer[2] data such as Visible Infrared Imaging Radiometer Suite (VIIRS) have become available but are still rarely used in the economics literature. According to Elvidge et al. (2013), VIIRS data have 45 times higher spatial resolution than DMSP data and have no blurring or geo-location errors (Gibson, 2021).

A few previous studies in economics have assessed the accuracy of night lights data in predicting GDP at various levels, such as national, regional and sub-regional levels (Chen and Nordhaus, 2011; Henderson et al. 2012; Keola et al. 2015; Hodler and Raschky, 2014) or to evaluate the accuracy of official national account statistics (Clark et al. 2017; Pinkovskiy and Martin, 2016). Some studies have found that nightlights can predict DHS wealth at local levels (Weidmann and Schutte, 2017) as well as human development outcomes created from DHS data using education, health and wealth (Bruederle and Hodler, 2018). However, these studies have all used DMSP data.

[2] Up until 2020, VIIRS offered only a monthly time-series. However, as of March 2021, an annual time series of VIIRS is also made available .

A parallel literature in remote sensing and artificial intelligence uses a combination of these alternative data sources and computer vision models to predict estimates of inequality or wealth. For example, Jean et al. (2016), trained a convolutional neural network (CNN) with high-resolution satellite imagery to predict local economic output at a local level for 5 countries using a transfer-learning approach. They did so by combining information from an image classification dataset trained to predict nightlights corresponding to daytime imagery and information from the Demographic and Health Surveys (DHS) and Living Standard Measurement Survey (LSMS). Since then, there have been several studies in this direction. Head et al. (2017) investigate whether the approach used by Jean et al. (2016) can be applied to other countries across other measures and find that even though the approach can be generalized across other countries, significant effort is required to fine-tune the hyperparameters to other geographic areas. They also find that the approach cannot be generalized to other measures of development such as access to drinking water among others. Kondmann and Zhu (2020) also adopt the same approach as Jean et al. (2016), to measure changes in poverty over time and find that the approach could not measure changes over time. More recently, Yeh et al. (2020), trained a CNN to predict wealth using daytime and nightlights imagery. They train separate models on each imagery and then combine them in a fully connected layer. Their findings suggest that models trained solely on nightlights or daytime imagery performed similarly and almost as well as the combined model, suggesting that the two inputs contain similar information. They conclude that their approach of using nightlights directly as inputs in their model performs better than Jean et al. (2016) that uses them indirectly.

Similarly, machine learning has been used with large scale network data such as CDR to get predictions of economic activity. For instance, Blumenstock et al. (2015) predict poverty at the individual level in Rwanda, by combining phone usage data with nightlight imagery and DHS survey data. Steele et al. (2017) also combines mobile operator aggregate data and satellite imagery to predict poverty measures for Bangladesh.

However, these methods don't come without limitations. According to a survey conducted by Asian Development Bank (ADB) and the United Nations Economic and Social Commission for Asia and the Pacific, incorporating big data into their programs is a challenge for 7 out of 16 National Statistical Offices (NSOs) in the ADB member countries due to difficulties in accessing these alternative data sources (Hofer et al. 2020). Moreover, while information from CDR can be helpful in generating poverty estimates when survey data is not available, it is limited by the fact that not all households own mobile phones which excludes certain sections of the population especially the poorest. In addition, mobile phone companies are hesitant in sharing data due to privacy and business concerns, which makes it challenging to rely on such data (Steele et al. 2017; Chinyama, 2022). Getting high-resolution satellite images is very expensive and that combined with the computer-intensive deep learning methods makes it difficult for them to be adopted widely especially by development organizations with limited resources. At the same time, these deep learning models are considered to be like a "black-box" in the sense that it is difficult for policy makers to understand and apply them to solving real world problems (Ledesma et al. 2020; Han et al. 2020). They are often criticized for attempting to only maximize model performance rather than model interpretability which is especially important when developing policies and interventions that have the potential of greatly affecting people's well-being (Ayush et al. 2020). For instance, a review done by Hall et al. (2023) of 32 papers on

wealth/poverty prediction that used satellite images as one of their inputs and deep neural networks as their method to predict survey data, finds that almost all literature conducted so far does not meet the requirements for interpretability and explainability, things that are important for wider acceptance/adoption of the research in the development community. They note that almost all research conducted has been by the "technical community" and therefore domain knowledge is mostly missing and should be integrated in future research. Further, model outputs from these studies are rarely made publicly available as rasters, dataframes or other data products that would be easy to incorporate into studies of the influence of poverty and household income by non-experts in remote sensing.

There are a handful of data sources on global variation in subnational economic activity that overcome these limitations, and are freely available for download by researchers . A recent study harmonized the two most widely used sources of night lights (DMSP and VIIRS), producing a new data source with global information on variation from 1992 to 2020 with high spatial resolution and that is directly comparable between different regions of the globe (Li et al. 2020).

As alternative sources for subnational wealth, Kummu et al. (2018) construct annual gridded datasets for GDP per capita (PPP) and Human Development Index (which measures achievements in important aspects of human development such as health, education and standard of living) from administrative records rather than household surveys. Finally, as an alternative to all of these sources, a primary source of subnational wealth is the information obtained directly from household surveys assessing variation in living conditions and household possessions are available for many years and countries from the DHS. In fact, most studies evaluating the accuracy of satellite data as proxies for economic activity do it against a household wealth index derived from DHS data.

Table 1: Data Description of Dependent and Independent Variables

Variable	Source	Spatial Coverage and Resolution	Temporal Coverage
Household wealth index	DHS	Over 90 countries, [GPS coordinates, displaced by up to 10 kms]	[since 1980s]
Nighttime lights data	Li et. al (2020)	Global,[30 arc-seconds]	[1992 to 2020]
Population density	GPWv4	Global, [30 arc-seconds]	[2000,2005,2010,2015,2020]
Human development index	Kummu et. al (2018)	39 countries, [5 arc-min]	[1990 to 2015]
Gross domestic product per capita	Kummu et. al (2018)	82 countries, [5 arc-min]	[1990 to 2015]

Though representing an important advance, these data sources vary widely in the extent of regions and years covered, as well as in their temporal and spatial resolution (Table 1), and so present important trade-offs for researchers to consider. There have also been relatively few studies directly comparing spatial patterns of variation in these indices. The use of night lights as a proxy of income variation has also been criticized because density of light sources is likely highly correlated with population density (Weidmann & Theunissen, 2021). Yet, few studies have yet assessed the degree to which the new harmonized nightlights data or alternative data sources (Table 1) vary independently of population density. Additionally, despite the increase in nighttime lights usage as a proxy of wealth, there are two significant knowledge gaps when it comes to the validity of using nighttime lights as a tool in social science research. Firstly, there is no research testing the accuracy of the new harmonized night lights data (Li et al. 2020) to measure economic activity and development at a small spatial scale, such as in developing country municipalities. Secondly, there are little to no studies (none except Bruederle and Hodler, 2018 and Head et al. 2017) on whether nighttime lights also indicate other important aspects of human development outcomes. Moreover, we are not aware of any research that studies the heterogeneity in the variation captured by nightlights (especially new harmonized nightlights) and population density separately for urban and rural areas. Taken together these issues present a complex and often confusing picture for researchers in areas such as economics, health sciences

and disease ecology that are not experts in GIS methods and the use of landsat data, but wish to incorporate information on spatial variation in poverty and economic activity into their work. In this paper, we fill these gaps.

Here we compare spatial variation in high quality indices of economic activity across 34 countries in sub-Saharan Africa. We focus on this region because it has often been considered one of the most challenging for which to obtain accurate information on subnational patterns of variation in socioeconomic factors, with many fewer data sources to choose from compared to better studied regions such as Europe and North America. It is also of inherent interest as an area of rapid population growth, swift ongoing changes in economic conditions (Akintunde and Oladeji, 2013) , and a region that presents a high risk of emergence of novel human diseases (Jones et al. 2008). We quantify the degree to which these indices (Table 1) explain or predict indicators of economic activity (wealth index, HDI and GDP per capita), and thus to which in any of them can be considered roughly equivalent sources of information. We also consider the degree to which Nightlights capture additional information beyond simple demographic variation (population density). Based on these comparisons, we provide recommendations for researchers that wish to incorporate poverty and other economic factors into spatially explicit modeling frameworks.

We find that even with the new harmonized nightlights dataset, overall nightlights are a good predictor of wealth index within countries at sub-national levels. We see that nightlights can explain variation in wealth even after controlling for population density. However, for our alternative dependent variables, such as HDI, we find that nightlights perform better at explaining variation in HDI across countries rather than within countries.

2. Data and Methods

2.1. DHS Wealth

The primary dependent variable in our study is derived from the DHS, which are a series of nationally representative and standardized household surveys conducted in low-income countries in Africa and other regions since the 1980s. These surveys provide information on health, fertility, education, wealth as well as many other socio-economic variables. Some of these surveys have geo-coordinates as well. The geographic information is however available at the level of survey clusters or Primary Sampling Units (PSUs) rather than at the level of households. A cluster is a group of 25-30 households that reside in close proximity to one-another. The clusters are categorized into urban and rural groups and the cluster location is reported with latitude/longitude coordinates (Weidmann & Theunissen, 2021). Most cluster locations are measured using GPS while only some are measured by information from gazetteers (Weidmann & Schutte, 2017). However, to ensure anonymity, cluster location points are randomly displaced by up to 2 kms for urban clusters and by 5 kms for rural clusters, with less than 1 percent of rural clusters displaced by up to 10 km (Bruederle and Hodler, 2018). Moreover, the DHS data are repeated cross-sections rather than a panel survey since a new sample of clusters is drawn for each round for the same country.

Our primary variable of interest is the DHS Wealth Score Index which is derived from data on household's ownership of a selected set of assets such as a phone, radio, car, TV, motorbike, etc; dwelling characterics like the number of rooms occupied in a home, flooring material, access to electricity, type of drinking water source, as well as other characteristics related to the wealth

status[3]. The DHS wealth is the most widely used proxy for poverty by previous machine learning literature on poverty estimation (Jean et al. 2016; Yeh et al. 2020). This is because the DHS collects detailed data on specific assets and uses a standardized method to compute a measure of wealth that is comparable across regions. However, due to reasons listed in the previous section, its coverage is patchy. Therefore, we use the DHS wealth score as the response variable that we are interested in predicting using proxies.

The DHS survey provides a household-level wealth index factor score which is calculated by analyzing the asset ownership of each individual through principal component analysis. It also provides a categorical household wealth index variable ranging from 1 to 5, derived from the household-level wealth factor score where 1 represents the lowest asset levels or "poorest" households and 5 represents the highest asset levels or "richest" households. However, as stated above, the DHS doesn't provide household locations but rather the average location of a group of households which is referred to as a household cluster. Therefore, we extract the Wealth index across all surveys in our sample and create average wealth index across all households within each cluster. For the wealth index factor variable, we normalize it so that it ranges from 0 to 1. This standardization process allows for easier comparison and analysis. To see the list of countries and years in our sample from DHS, see Table A.1 in Appendix.

[3] In particular, we use the Household Recode survey data for each country (or HR) which contains all the attributes of each household and the corresponding GPS dataset for the same year and phase. For details on how DHS wealth index is constructed see Rutstein and Johnson, K. (2004).

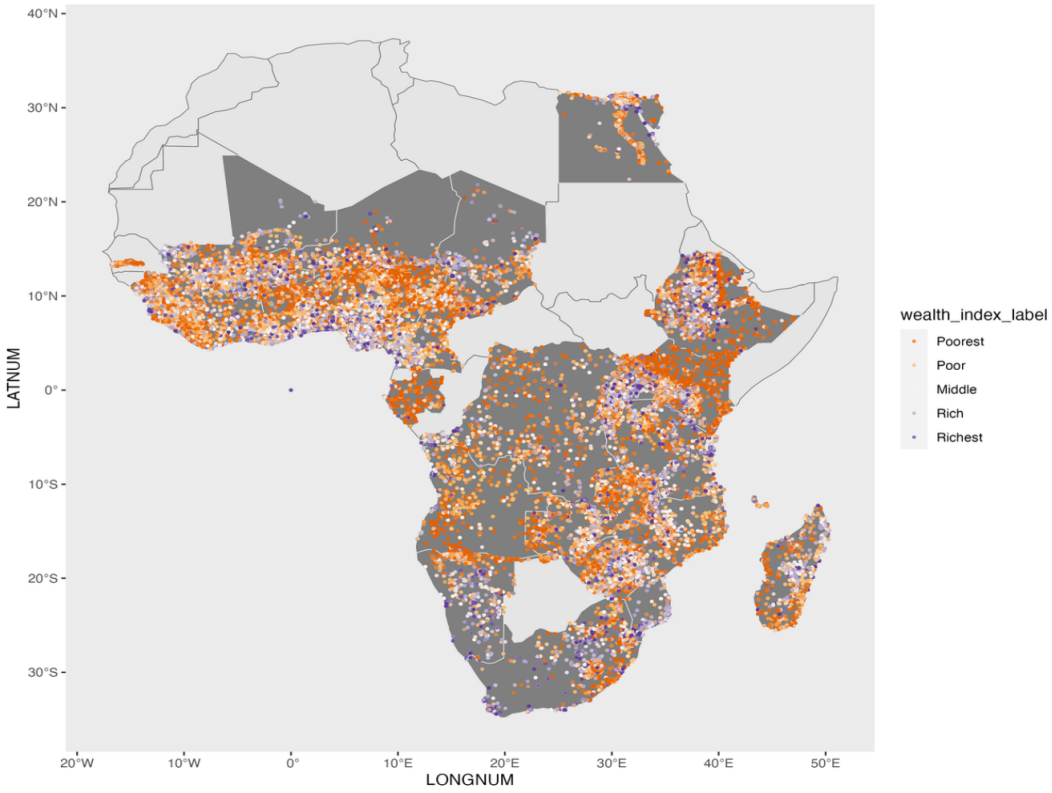


Figure 1: Geographic distribution of DHS survey clusters in Africa. The map shows the location of survey clusters with each point representing a cluster. The countries shaded in "gray" color are the countries in our sample. The colors of the points indicate the wealth category of the households surveyed.**Note:** The DHS wealth index is a relative measure of wealth and the methodology used to compute it does not allow for cross-country comparisons.

In figure 1 below we map the geographic location of the DHS clusters in our sample and color those clusters according to their wealth index category . Note that due to the methodology used to construct the wealth index, it should only be compared within a country.

2.1. Nightlights data

We use a harmonized nightlights dataset generated by Li et al. 2020. This gives us the opportunity to expand the time-period for our study since data from the two sources DMSP (available for 1992-2013) and VIIRS (available for 2013-2019) are not comparable over time because of little temporal overlap and also because of different spatial resolution.

Our dataset uses nightlights data from DMSP and VIIRS to produce a harmonized and consistent series from 1992 to 2020. The DMSP nightlights data is a time series dataset with a spatial resolution of 30 arc-seconds covering 180°W to 180°E in longitude and 65°S to 75°N in latitude. However, due to factors such as varied atmospheric conditions, satellite shift, and sensor degradation, the raw DMSP nightlights data is not comparable across years. To overcome this limitation, a stepwise calibration approach was used to generate a temporally consistent nightlights dataset, which outperforms traditional approaches in terms of temporal trends and correlation with electricity consumption data. In contrast, the VIIRS Day/Night Band (DNB) data provides radiance records with improved radiometric resolution and a higher spatial resolution of 15 arc-seconds. The VIIRS product is a suite of global average radiance composite images that have been corrected for effects caused by biogeophysical processes and stray light. The monthly VIIRS nightlights data were further preprocessed and composited as annual time series data. The dataset covers the latitudinal zone of 65°S-75°N and spans from 2012 to 2020.

To generate the harmonized series from 1992 to 2020, Li et al. (2020) used a three-step framework. Firstly, annual VIIRS nightlights data was produced from monthly observations and noise from temporary lights such as fires and boats was excluded. Secondly, the relationship between processed VIIRS data and DMSP nightlights data in 2013 was quantified using a sigmoid function. Thirdly, the derived relationship was applied globally to obtain DMSP-like data from VIIRS. The consistent nightlights data was consistent nightlights data was generated by integrating the temporally calibrated DMSP nightlights data from 1992 to 2013 and DMSP-like nightlights data from VIIRS from 2014 to 2020 (Li et al. 2020). To see how nightlights correlate with Wealth, see scatterplots for six countries in Appendix.

The dataset from Li et al. (2020) is available yearly, downloadable as .tif files and at a spatial resolution of 30 arc-seconds (~1 km). We aggregate them at 10 km^2 to match them with Demographic and Health Survey (DHS) household wealth data using the geographic coordinates at the cluster level in the DHS surveys.

2.2. Population density data

We use population density estimates from the CIESIN Gridded Population of the World (GPWv4). These data were also downloaded in GeoTIFF format at a spatial resolution of 30 arc-second or (~1 km) for the years 2000, 2005, 2010, 2015 and 2020 and for the years between these intervals, we estimate population density using linear interpolation. Similar to nightlights, these were then also aggregated to 10 km^2 and combined with DHS data for the years 2004 to 2019. To see how population density correlates with Wealth, see scatterplots for six countries in Appendix.

2.3. Alternative dependent variables: (HDI and GDP per capita)

Along with the DHS wealth index, we also use two alternative dependent variables: theHuman Development Index and Gross Domestic Product (per capita). Both these datasets were obtained from the Kummu et al. (2018).

Kummu et al. use both national-level data (from UNDP) and subnational-level from various sources such as censuses and UNDP reports for areas outside of Europe, and Eurostat for Europe) to construct HDI at 5 arc-min level of resolution. On the other hand, they constructed GDP per capita dataset at 5 arc-min resolution using national GDP per capita (from the World

Bank dataset and CIA's World Factbook) and the subnational GDP per capita data (PPP) based on Gennaioli et al. (2013).

Below, we plot two maps showing the variation in the HDI and GDP per capita dataset for the DHS cluster locations. We divide the data into 5 bins to plot them. We note that there is very little variation within countries for both of these datasets. This goes against the motivation of these datasets which were created to fill in the gap for the need for datasets that explain sub-national variation. While their study reports sub-national variation for larger countries such as the United States, China, India etc.,we don't notice any such variation for the sample of 34 countries in our sample in Africa.

Figure 2: Geolocated DHS clusters in Africa, colored by HDI

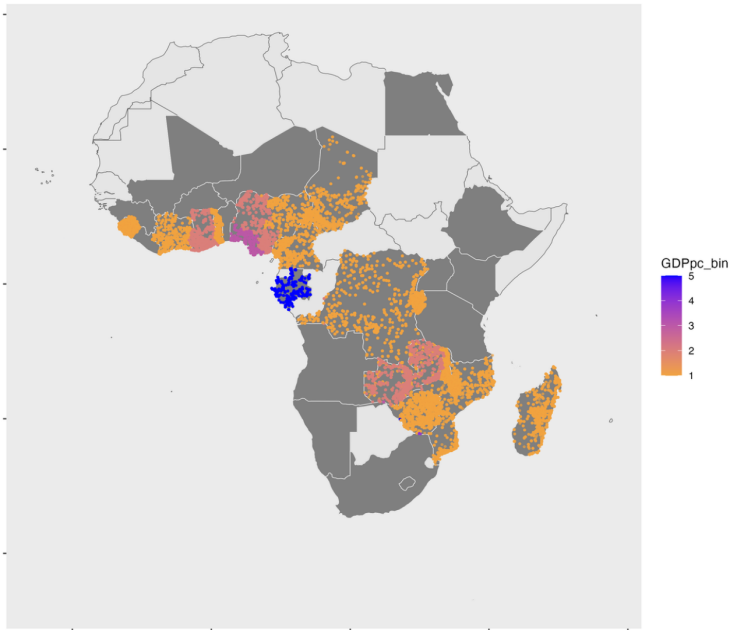
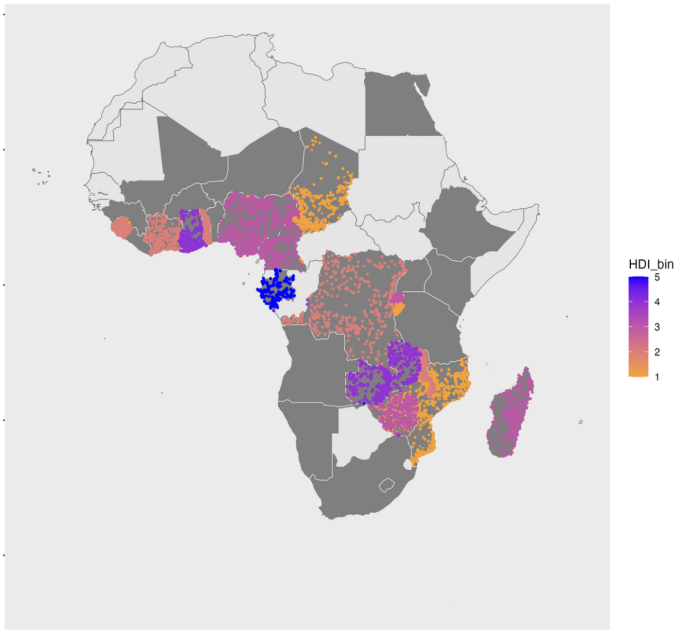


Figure 3: Geolocated DHS clusters in Africa, colored by HDI



Descriptive Statistics

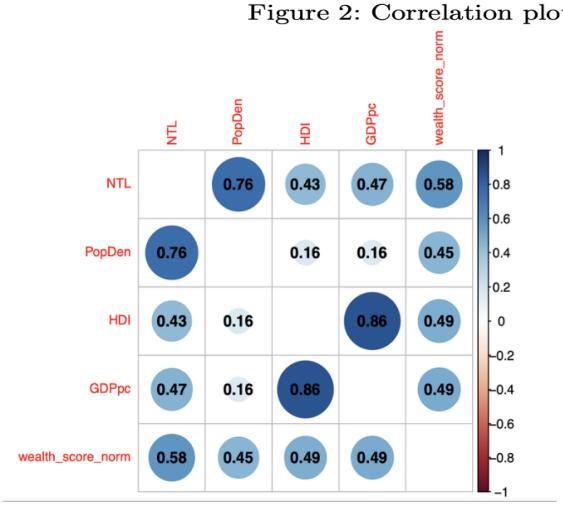
We present descriptive statistics for our variables across urban and rural samples in Table 2 below. We see that, the mean of log nightlights is -1.6 in rural areas, which shows that on average, rural areas have a lower level of light emissions at night compared to urban areas (2.1). The standard deviation of 3.2 indicates that there is a higher degree of variability in the level of nightlights across different rural areas compared to 2.1 in urban areas. Similarly, we also see that the mean of log population is 4.7 in rural areas, which shows that on average, rural areas have a smaller population density compared to urban areas (6.5). The standard deviation of 1.7 indicates that there is a considerable amount of variation in population density across different rural areas, although it is relatively smaller than the variability in urban areas.

Table 2: Summary Statistics for our variables in Urban versus Rural Sample

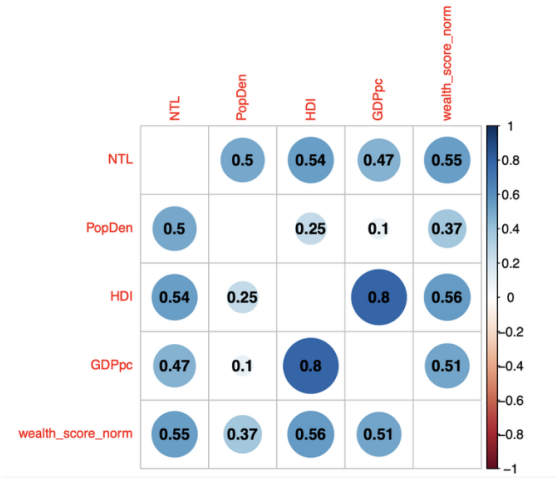
	Rural						Urban					
Variable	N	Mean	SD	Min	Max	Median	N	Mean	SD	Min	Max	Median
Mean wealth Index	24794	2.4	0.81	1	5	2.3	14278	4.1	0.86	1	5	4.4
Log(Nightlights)	24794	-1.6	3.2	-4.6	4.1	-4.6	14278	2.1	2.2	-4.6	4.1	2.6
Log(Pop Den)	24794	4.7	1.7	-4.6	11	4.7	14278	6.5	2.1	-4.6	11	6.5
Log(Human Development Index)	19720	-0.74	0.18	-1.2	-0.36	-0.74	11160	-0.68	0.19	-1.2	-0.37	-0.69
Log(GDP per capita)	19720	7.7	0.83	5.9	9.8	7.4	11160	8	0.9	5.9	9.8	7.9

As a first step, we calculate the Spearman correlation coefficient matrix for the variables in our data separately for urban and rural areas (Figure 2 below). We see that Nightlights and Population density are positively and strongly correlated (0.76) in the urban sample as well as the rural sample (0.5). Surprisingly, nightlights have the same level of correlation with wealth in both urban and rural sample. Population density, on the other hand, has a higher correlation with wealth in the urban sample than in the rural sample. Wealth and other alternative indicators of wealth such as HDI and GDP per capita are positively correlated with DHS wealth in both urban and rural samples. Interestingly, Population density has a low correlation with both HDI and GDP per capita.

Figure 2: Correlation plots for urban and rural areas.



(a) Urban correlation plot



(b) Rural correlation plot

Empirical Specification

We estimate the following relationship using simple linear regressions:

$$(Eq.1)\quad Wealth_{(cjt)} = \beta_0 + \beta_1\, ln(light_{(cjt)}) + \beta_2\, ln(population_{(cjt)}) + \beta_3 (D)_{(cjt)} + \beta_4 (D)_{(cjt)} * ln(light_{(cjt)}) + \beta_5 (CC)*ln(lights_{(cjt)}) + \beta_6 (CC)*ln(population_{(cjt)}) + \Gamma_{(jt)} + e_{(jt)}$$

where, $Wealth_{cjt}$ is wealth measured by our 3 dependent variables: 1) Our main dependent variable - DHS Wealth index that captures household wealth in cluster c, country j and year t; 2) Human Development Index; and 3) GDP per capita. Variables 2) and 3) are first obtained from Kummu et al. (2018) and then extracted at the DHS cluster locations. Next, our right hand side variables include: $light_{cjt}$ which is nighttime light intensity; $population_{cjt}$ which is population density ; $(D)_{cjt}$ is a dummy variable that equals 1 if the cluster is in Urban area and 0 otherwise, Γ_{jt} are country and year fixed effect, and e_{jt} is the error term. For our $light_{cjt}$ and $population_{cjt}$ variables, we add a small constant (0.01) to them, before taking a log (following Bruederle and

Hodler, 2018). The $(D)_{cjt}$ dummy for Urban area will control for any effect due to the variation between urban and rural areas. In some columns, we also add an interaction term between urban dummy and nightlight intensity $(D)_{(cjt)} * ln(light_{(cjt)})$, to measure the difference in the slope (of nightlights) effects between the two groups. It captures the differential between Urban versus Rural areas across levels of light intensity.

We add country and year fixed effects in all regressions with DHS wealth index as the dependent variable since wealth index is purposely constructed differently for each county. Further, adding country and year fixed effects will allow us to examine the relationship of our proxies with wealth within country and year groups. Country fixed effects control for country-specific characteristics that do not vary over time (i.e. help focus on the within country variation). Year dummies control for global shocks (e.g. technological changes) that affect all countries. In some columns, we also add a) interaction between nightlights and country dummy $(CC)*ln(lights_{(cjt)})$ which captures the idea that the relationship between nightlights and wealth may be different in different countries, and/or b) interaction between population and country dummy $(CC)*ln(population_{(cjt)})$ which captures a similar idea. We also cluster standard errors at the same level of fixed effects (country and year) which will adjust for potential correlations between error terms within the same country or year.

3. Results

Table 3: OLS regression results with country/year fixed effects. Standard errors clustered by country and year

	Wealth Index				
	mean_WI				
	(1)	(2)	(3)	(4)	(5)
Log Nightlights	0.073***	0.208***	0.087***	0.135***	
	(0.008)	(0.008)	(0.002)	(0.002)	
Log Population Density	0.108***	0.200***	0.094***		0.178***
	(0.012)	(0.004)	(0.004)		(0.003)
Urban (=1)	1.206***		1.210***	1.367***	1.494***
	(0.012)		(0.012)	(0.011)	(0.011)
Urban.Nightlights	0.068***		0.072***		
	(0.004)		(0.004)		
Constant	1.343***	1.614***	1.374***	1.713***	1.093***
	(0.051)	(0.033)	(0.026)	(0.026)	(0.027)
Fixed effects (Country and year)	Yes	Yes	Yes	Yes	Yes
Interaction (Log Nightlights:Country)	Yes	Yes	No	No	No
Interaction (Log Population:Country)	Yes	No	No	No	No
Observations	39,072	39,072	39,072	39,072	39,072
R^2	0.649	0.482	0.640	0.619	0.611
Adjusted R^2	0.648	0.481	0.640	0.618	0.610
Residual Std. Error	0.695 (df = 38963)	0.844 (df = 38965)	0.703 (df = 38996)	0.724 (df = 38998)	0.731 (df = 38998)
F Statistic	665.986*** (df = 108; 38963)	341.996*** (df = 106; 38965)	925.628*** (df = 75; 38996)	867.272*** (df = 73; 38998)	838.703*** (df = 73; 38998)
Note:					*p<0.1; **p<0.05; ***p<0.01

Tables 3 and 4, show regressions estimated using DHS wealth index as the dependent variable, table 5 and 6 use HDI as the dependent variable.

In table 3 above, Nightlights alone can explain 61 percent of the variation in wealth. For all columns (1 to 5) the coefficient on nightlights is positive and significant at 1 percent, indicating that higher levels of nighttime light emissions are associated with higher mean household wealth, all else being equal. The coefficient on population density is also positive and significant across all columns, indicating that larger population sizes are associated with higher mean household wealth, all else being equal. The coefficients of Urban in columns 1, 3, 4 and 5 are also positive and significant, indicating that urban areas have higher mean household wealth than rural areas.

The positive and significant coefficient on Urban * Nightlights indicate that the positive relationship between nighttime light emissions and mean household wealth is stronger in urban areas compared to rural areas. This suggests that nighttime light emissions are a more accurate proxy for economic activity in urban areas.

Moving from column 3 to column 1, we see that adding the interaction terms-nightlights:country and population density:country, the R square doesn't change by much, indicating that these variables do not improve the model's ability to predict wealth, however, the slope coefficients still remain significant even though there is a reduction in the size of the slope coefficients.

Table 4: Predictive Power of Nightlights versus Population Density for Wealth Index in Urban/ Rural Areas

	Mean Wealth Index							
	mean_WI							
	All-NTL	Urban-NTL	Rural-NTL	All-PopD	Urban-PopD	Rural-PopD	Urban-NTL-PopD	Rural-NTL-PopD
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Log Nightlights	0.282***	0.311***	0.082***				0.160***	0.097***
	(0.009)	(0.034)	(0.012)				(0.005)	(0.003)
Log Population Density				0.327***	0.209***	0.107***	0.079***	0.112***
				(0.013)	(0.013)	(0.034)	(0.005)	(0.005)
Constant	2.314***	2.540***	1.721***	1.339***	2.424***	1.323***	2.590***	1.411***
	(0.031)	(0.107)	(0.040)	(0.066)	(0.079)	(0.095)	(0.038)	(0.035)
Fixed effects (Country and year)	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Interaction (Log Nightlights:Country)	*Yes*	*Yes*	*Yes*	*No*	*No*	*No*	*No*	*No*
Interaction (Log Population Density:Country)	*No*	*No*	*No*	*Yes*	*Yes*	*Yes*	*No*	*No*
Observations	39,072	14,278	24,794	39,072	14,278	24,794	14,278	24,794
R^2	0.436	0.435	0.224	0.394	0.405	0.247	0.426	0.218
Adjusted R^2	0.434	0.430	0.220	0.392	0.401	0.243	0.423	0.215
Residual Std. Error	0.881	0.646	0.712	0.913	0.662	0.701	0.650	0.714

Note: *p<0.1; **p<0.05; ***p<0.01

Table 4 above shows the results for assessing the predictive power of nightlights and population density for wealth index in urban versus rural areas. Consistent with the results in Table 3, in urban areas, nightlights explain approximately two times the share of variation (43%) in wealth as compared to rural areas (22 %). Similarly, the prediction power of population density is much higher for urban areas (40 %) as compared to rural (24 %). In column (1), the coefficient on Nightlights indicates that a 100 percent increase in nightlights is associated with a 0.282-unit increase in household mean wealth, holding other variables constant. Similarly, in column (4), the coefficient on Population Density indicates that a 100 percent increase in population density is associated with a household mean wealth.

Table 5: OLS regression results with country/year fixed effects. Standard errors clustered by country and year

	log (Human Development Index)				
	(1)	(2)	(3)	(4)	(5)
Log Nightlights	0.0003***	0.00004			0.0002
	(0.0001)	(0.00003)			(0.0002)
Log Population Density			−0.00003	−0.0002	−0.0004
			(0.0001)	(0.0003)	(0.0004)
Urban (=1)					0.0001
					(0.0004)
Constant	−0.629***	−0.629***	−0.629***	−0.628***	−0.627***
	(0.0004)	(0.0004)	(0.001)	(0.001)	(0.002)
Fixed effects (Country and year)	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Interaction (Log Nightlights:Country)	*No*	*Yes*	*No*	*No*	*Yes*
Interaction (Log Population Density:Country)	*No*	*No*	*No*	*Yes*	*Yes*
Observations	30,880	30,880	30,880	30,880	30,880
R^2	0.982	0.982	0.982	0.982	0.983
Adjusted R^2	0.982	0.982	0.982	0.982	0.983
Residual Std. Error	0.025 (df = 30822)	0.025 (df = 30791)	0.025 (df = 30822)	0.025 (df = 30791)	0.025 (df = 30758)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5 and 6 estimates equation 1 using the Human development index as the dependent variable. Here, we have log of HDI as the dependent variable for ease of interpretation of the slope coefficients in terms of elasticities. In Table 5, the coefficients in column 1, suggest a positive and significant relationship between HDI and nightlights, after controlling for country and year fixed effects. The high R-squared value of 98% suggests that the model explains a large proportion of the variation in HDI. However, in column 2, when adding the interaction term (Nightlights * Country) to capture the idea that the effect of nightlights on HDI may differ across countries , the coefficient on nightlights becomes insignificant. In columns 3, 4, and 5, we see no statistically significant relationship between population and HDI. The intercept is significant, even though the slope coefficients are not, indicating that the only variation we see is from country and year dummy.

In table 5, we looked at variation in HDI explained by nightlights and population density within countries and years by using country and year fixed effects. Now, in Table 6, we repeat those regressions using just time fixed effects to look at across country variation in HDI. Here, we see clearly that nightlights alone can explain upto 65 percent of the variation in HDI, across countries. While population density also, alone can explain up to 84 percent variation in HDI

across countries. We see that overall, both nightlights and population density perform better at predicting HDI across countries rather than within country variation. This result can also be seen from figure 3.

Table 6: OLS regression results with year fixed effects. Standard errors clustered by year

	log (Human Development Index)				
	(1)	(2)	(3)	(4)	(5)
Log Nightlights	0.021***	0.005***			0.003***
	(0.0003)	(0.001)			(0.0003)
Log Population Density			0.015***	−0.004***	−0.011***
			(0.001)	(0.001)	(0.001)
Urban (=1)					−0.003***
					(0.001)
Constant	−0.829***	−0.874***	−0.948***	−0.836***	−0.794***
	(0.003)	(0.003)	(0.004)	(0.003)	(0.005)
Fixed effects (year)	*Yes*	*Yes*	*Yes*	*Yes*	*Yes*
Interaction (Log Nightlights:Country)	*No*	*Yes*	*No*	*No*	*Yes*
Interaction (Log Population Density:Country)	*No*	*No*	*No*	*Yes*	*Yes*
Observations	30,880	30,880	30,880	30,880	30,880
R^2	0.387	0.652	0.286	0.849	0.922
Adjusted R^2	0.387	0.652	0.286	0.849	0.922
Residual Std. Error	0.148 (df = 30867)	0.111 (df = 30836)	0.159 (df = 30867)	0.073 (df = 30836)	0.053 (df = 30803)

Note: *p<0.1; **p<0.05; ***p<0.01

In the future we also plan to test the above relationship with GDP per capita from Kummu et al. 2018. Also further, as robustness checks we will use the median DHS wealth index, international wealth index and mean DHS wealth score as dependent variables.

Conclusion

This study aimed to address the challenges associated with obtaining accurate subnational data on social, economic, and political realities in developing countries. The limited availability and quality of subnational data in these regions have led researchers to explore alternative data sources such as night lights, satellite imagery, and mobile phone records. By comparing spatial variation in economic activity indices across 34 sub-Saharan African countries, this research has provided valuable insights into the validity and usefulness of these alternative data sources.

The findings demonstrate that night lights, particularly the new harmonized dataset, serve as a reliable proxy for wealth index within countries at subnational levels, even after controlling for population density. This suggests that night lights can effectively capture variations in economic activity and provide valuable information for understanding local-level socioeconomic dynamics. However, it should be noted that night lights perform better at explaining variation in broader indicators like the Human Development Index (HDI) across countries rather than within countries. This indicates that while night lights offer valuable insights into economic activity, they may have limitations in capturing other dimensions of human development outcomes.

The recommendations provided based on the analysis of these alternative data sources can guide researchers in selecting appropriate indicators and incorporating them into spatially explicit modeling frameworks.

References

- Akintunde, T. S., & Oladeji, P. A. O. S. I. (2013). Population dynamics and economic growth in Sub-Saharan Africa. *Population*, *4*(13).
- Ayush, K., Uzkent, B., Burke, M., Lobell, D., & Ermon, S. (2020). Generating interpretable poverty maps using object detection in satellite images. *arXiv preprint arXiv:2002.01612*.
- Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073-1076.
- Blumenstock, J. E. (2016). Fighting poverty with data. *Science*,*353*(6301), 753–754. https://doi.org/10.1126/science.aah5217

- Bruederle, A., & Hodler, R. (2018). Nighttime lights as aproxy for human development at the local level. *PloS one*, *13*(9), e0202231.
- Center For International Earth Science Information Network-CIESIN-Columbia University.(2016). *Gridded Population of the World, Version 4 (GPWv4): Population Count* [Data set]. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). https://doi.org/10.7927/H4X63JVC
- Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences*, *108*(21), 8589-8594.
- Chinyama, Francis. "Predicting household poverty with machine learning methods: the case of Malawi." Master's thesis, Faculty of Science, 2022.
- Clark, H., Pinkovskiy, M., & Sala-i-Martin, X. (2017). *China's GDP growth may be understated* (No. w23323). National Bureau of Economic Research.
- Gibson, J., Olivia, S., & Boe-Gibson, G. (2020). Night Lights in Economics: Sources and Uses1. *Journal of Economic Surveys*, *34*(5),955–980. https://doi.org/10.1111/joes.12387
- Gibson,J., Olivia, S., Boe-Gibson, G., & Li, C. (2021). Which night lights data should we use in economics, and where? *Journal of Development Economics*, *149*, 102602. https://doi.org/10.1016/j.jdeveco.2020.102602
- Gibson, J. (2021). Better night lights data, for longer. *Oxford Bulletin of Economics and Statistics*, *83*(3), 770-791.
- Goldblatt, R., Heilmann, K., & Vaizman, Y. (2020). Can medium-resolution satellite imagery measure economic activity at small geographies? Evidence from Landsat in Vietnam. *The World Bank Economic Review*, *34*(3), 635-653.

- *Harmonization of DMSP and VIIRS nighttime light data from 1992-2020 at the global scale*. (2020). [Data set]. figshare. https://doi.org/10.6084/m9.figshare.9828827.v5
- Han, S., Ahn, D., Park, S., Yang, J., Lee, S., Kim, J., ... & Cha, M. (2020, August). Learning to score economic development from satellite imagery. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2970-2979).
- Hall, O., Dompae, F., Wahab, I., & Dzanku, F. M. (2023). A review of machine learning and satellite imagery for poverty prediction: Implications for development research and applications. *Journal of International Development*.
- Head, A., Manguin, M., Tran, N., & Blumenstock, J. E. (2017). Can human development be measured with satellite imagery?. *Ictd*, *17*, 16-19.
- Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American economic review*, *102*(2), 994-1028.
- Hodler, R., & Raschky, P. A. (2014). Regional favoritism. *The Quarterly Journal of Economics*, *129*(2), 995-1033.
- Hofer, M., Sako, T., Martinez Jr, A., Addawe, M., Bulan, J., Durante, R. L., & Martillan, M. (2020). Applying Artificial Intelligence on Satellite Imagery to Compile Granular Poverty Statistics. *Asian Development Bank Economics Working Paper Series*, (629).
- Jean, N., Burke, M., Xie, M., Davis, W. M., Lobell, D. B., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790-794.

- Jones, K. E., Patel, N. G., Levy, M. A., Storeygard, A., Balk, D., Gittleman, J. L., & Daszak, P. (2008). Global trends in emerging infectious diseases. *Nature*, *451*(7181), 990-993.
- Keola, S., Andersson, M., & Hall, O. (2015). Monitoring economic development from space: using nighttime light and land cover data to measure economic growth. *World Development*, *66*, 322-334.
- Kondmann, L., & Zhu, X. X. (2020). Measuring changes in poverty with deep learning and satellite imagery.
- Kummu, M., Taka, M., & Guillaume, J. H. (2018). Gridded global datasets for gross domestic product and Human Development Index over 1990–2015. *Scientific data*, *5*(1), 1-15.
- Ledesma, C., Garonita, O. L., Flores, L. J., Tingzon, I., & Dalisay, D. (2020). Interpretable poverty mapping using social media data, satellite images, and geospatial information. *arXiv preprint arXiv:2011.13563*.
- Li, X., Zhou, Y., Zhao, M., & Zhao, X. (2020). A harmonized global nighttime light dataset 1992–2018. *Scientific data*, *7*(1), 168.
- Pinkovskiy, M., & Sala-i-Martin, X. (2016). Lights, camera… income! Illuminating the national accounts-household surveys debate. *The Quarterly Journal of Economics*, *131*(2), 579-631.
- Rutstein, S. O., & Johnson, K. (2004). *The DHS wealth index* (No. 6). ORC Macro, MEASURE DHS.
- Sanghi, A., Bundervoet, T., & Maiyo, L. (2015). Night lights and the pursuit of subnational GDP: Application to Kenya & Rwanda. *Banque Mondiale. Repéré à*

http://blogs.worldbank.org/developmenttalk/night-lights-and-pursuit-subnational-gdp-application-kenya-rwanda.

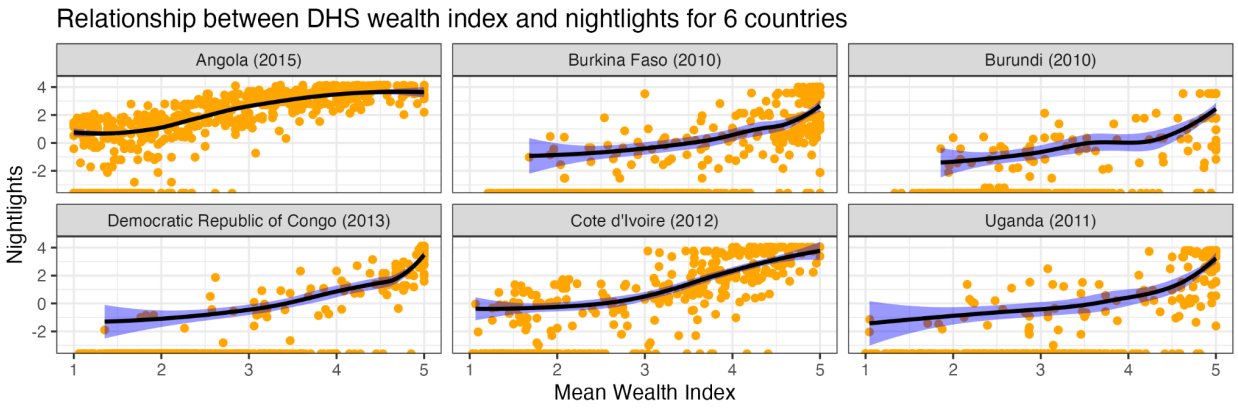
- Steele, J. E., Sundsøy, P. R., Pezzulo, C., Alegana, V. A., Bird, T. J., Blumenstock, J., ... & Bengtsson, L. (2017). Mapping poverty using mobile phone and satellite data. *Journal of The Royal Society Interface*, *14*(127), 20160690.
- Steele, J. E., Pezzulo, C., Albert, M., Brooks, C. J., zu Erbach-Schoenberg, E., O'Connor, S. B., ... & Tatem, A. J. (2021). Mobility and phone call behavior explain patterns in poverty at high-resolution across multiple settings. *Humanities and Social Sciences Communications*, *8*(1), 1-12.
- Weidmann, N. B., & Schutte, S. (2017). Using night light emissions for the prediction of local wealth. *Journal of Peace Research*, *54*(2), 125–140. https://doi.org/10.1177/0022343316630359
- Weidmann, N. B., & Theunissen, G. (2021). Estimating Local Inequality from Nighttime Lights. *Remote Sensing*, *13*(22), Article 22. https://doi.org/10.3390/rs13224624
- Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., & Burke, M. (2020). Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-16185-w

Appendix

Table A.1: List of countries and DHS waves in our sample

Country	DHS survey years
Angola	2015
Benin	2012,2017
Burkina Faso	2010
Burundi	2010,2016
Cameroon	2004, 2011, 2018
Chad	2014
Comoros	2012
Congo Democratic Republic	2007,2013
Cote d'Ivoire	2012
Egypt	2005,2008,2014
Eswatini	2006
Ethiopia	2005,2010,2016,2019
Gabon	2012
Gambia	2019
Ghana	2008,2014
Guinea	2005,2012,2018
Kenya	2008,2014
Lesotho	2004,2009,2014
Liberia	2007,2013,2019
Madagascar	2008
Malawi	2004,2010,2015
Mali	2006,2012,2018,
Mozambique	2011
Namibia	2006,2013
Niger	2012
Nigeria	2008,2013,2018
Rwanda	2005,2008,2010,2014,2019
Sierra Leone	2008,2013,2019
South Africa	2017
Tanzania	2010,2015
Togo	2013
Uganda	2006,2011,2016
Zambia	2007,2013,2018
Zimbabwe	2005, 2010, 2015

The scatterplots below show the relationship between log nightlights and wealth index at the cluster level for 6 countries in our sample (other countries in the appendix). The black lines show the LOESS fit which captures the general trend in the data. The blue shaded area shows the 95 percent confidence interval. From these plots we see that in general,the locations with low nightlights are also the locations with the lowest wealth.



The scatterplots below show the relationship between log population density and wealth index at the cluster level for 6 countries in our sample. We see that in general, the locations with low population density are also the locations with the lowest wealth.