Economic Development in Pixels: The Limitations of Nightlights and New Spatially Disaggregated Measures of Consumption and Poverty

BSE Working Paper 1433| February 2024

John D. Huber, Laura Mayoral

bse.eu/research

Economic development in pixels: The limitations of nightlights and new spatially disaggregated measures of consumption and poverty[1]

John D. Huber
Columbia University
Laura Mayoral
Institute for Economic Analysis and Barcelona School of Economics

February 2024

Abstract

We develop a novel methodology that uses machine learning to produce accurate estimates of consumption per capita and poverty in 10x10km cells in sub-Saharan Africa over time. Using the new data, we revisit two prominent papers that examine the effect of institutions on economic development, both of which use "nightlights" as a proxy for development. The conclusions from these papers are reversed when we substitute the new consumption data for nightlights. We argue that the different conclusions about institutions are due to a previously unrecognized problem that is endemic when nightlights are used as a proxy for spatial economic well-being: nightlights suffer from nonclassical measurement error. This error will typically lead to biased estimates in standard statistical models that use nightlights as a spatially disaggregated measure of economic development. The bias can be either positive or negative, and it can appear when nightlights are used as either a dependent or an independent variable. Our research therefore underscores an important limitation in the use of nightlights, which has become the standard measure of spatial economic well-being for studies focusing on developing parts of the world. It also demonstrates how machine learning models can generate a useful alternative to nightlights, with important implications for the conclusions we draw from the analyses in which such data are employed.

Keywords: Economic development, poverty, institutions, nightlights, nonclassical measurement error, machine learning.

JEL codes: C01, P46, P48

[1]We are grateful for comments from Debraj Ray and seminar participants at Columbia, Kent University, WZB Berlin and UC-Berkeley, and for research assistance from Ella Bayi, Martin Devaux, Dylan Groves, Salif Jaiteh, Alex Nino, and Adriana Oliveras. Huber gratefully acknowledges Columbia University for use of the High Performance Computing cluster. Mayoral gratefully acknowledges financial support through grant PGC2018-096133-B-100 and PID2021-124256OB-I00 funded by MCIN/AEI/10.13039/501100011033 and by ERDF "A way of making Europe", Severo Ochoa Program for Centers of Excellence (CEX2019-000915-S), AGAUR-Generalitat de Catalunya (2021 -SGR-416) and La Caixa Foundation Research Grants on Socio-Economic Well-Being through the project "Inequality, Political Instability and Long-term Development". First version: June 2023.

1. Introduction

Spatially disaggregated data on key measures of economic development are crucial to the study of a wide-range of questions about economic progress, violence and conflict, and policies to alleviate poverty, among others. Such data, however, is lacking for much of the developing world, and especially in Africa. Scholars have therefore turned to the use of satellite images of luminosity at night ("nightlights"). Pioneered by Henderson et al (2011, 2012) and Chen and Nordhaus (2011), nightlights are a natural proxy for economic development, with brighter areas associated with higher development. Since nightlights are measured in very small pixels (grid cells are measured at a resolution of 30 arc-seconds, or in cells that are approximately 1 km^2 at the equator), they can be aggregated at essentially any spatial level. And though changes over time in satellite technology create challenges, time series analysis is possible because yearly data exist for the whole world since 1992. Given these features of nightlights and the paucity of alternatives, it is not surprising that so many scholars now use nightlights in empirical research on economic development and to evaluate economic outcomes.[2]

However, there are two important limitations associated with using nightlights (NL) as a spatially fine-grained proxy for development. The first is well-known: nightlights do not have a substantively interpretable metric, providing at most ordinal information about spatial economic activity. But measures expressed in widely understood cardinal metrics, like income or consumption levels, are essential because they make it possible to calculate poverty rates, growth rates, levels of inequality, and other key indicators of spatial economic development.

The second limitation has received little if any attention in the previous literature: nightlights suffer measurement error that is typically non-classical. That nightlights suffer measurement error has of course been widely recognized, including by Chen and Nordhaus (2011).[3] But the fact that measurement error is non-classical poses a considerable challenge for research using NL as a proxy for spatially disaggregated development in regression models because estimates from such models will typically be biased.

[2]See Michalopoulos and Papaioannou (2018), and Gibson, Olivia and Boe-Gibson (2020) for reviews of the literature in economics, and see Burke et. al. (2021) for a discussion of how satellite imagery has been used for assessing progress toward sustainable development goals. Figure A.1 in Appendix A provides evidence of the rapid increase in the use of nightlights in economics research over time.

[3]See Gibson, Olivia and Boe-Gibson (2020) for discussions of the sources of measurement error.

This paper makes two contributions. First, we develop new spatially disaggregated estimates of economic well-being that are denominated in an easily interpretable metric, consumption dollars. Second, we use these data to undergird our argument that measurement error in NL is non-classical, leading to biases of unknown direction in estimates from regression models that use NL as a dependent or independent variable.

To develop spatially disaggregated estimates of economic well-being, our approach follows a strategy used in previous research.[4] In broad strokes, this strategy uses individual-level surveys that have information on asset ownership to create an asset index that serves as a proxy for economic well-being. Such surveys are obviously very limited in their geographic scope (particularly in developing parts of the world), and they are conducted infrequently. But the asset indices are geocoded, and thus can be used to create a training variable for use in machine learning models. The asset-based training variable can be predicted by variables like nightlights, making it possible to estimate economic well-being across essentially any space and over time. The results of such efforts can therefore be used to assess development outcomes, as summarized in Burke et. al (2021).

A central limitation of these existing studies, however, is that the asset index used as a training variable lacks interpretability (Yeh et al. 2020, McCallum et al., 2022) and sometimes even cross-country comparability (Chi et al. 2022).[5] In this sense, it shares the same limitation as nightlights. Our approach addresses this limitation by developing a training variable that measures the log of consumption per capita in 2011 PPP dollars. We do this by developing a simple mathematical framework for combining individual-level asset indices with macro data on consumption per capita and inequality to create individual-level measures of log consumption per capita.

More specifically, we generate geographically fine-grained measures of consumption in three steps. First, we construct the new training variable, denominated in (log) consumption dollars, using geocoded household-level information on asset ownership from the Demographic and Health Surveys (DHS), along with the World Bank's WB-PIP data on

[4] See, e.g., Jean et. al. (2016), Yeh et al. (2020 and 2021), Chi et al. (2022) and Aiken et al. (2022).

[5] An additional limitation is that to create an asset-based training variable that is comparable across time and space requires the strong assumption that the economic value of particular assets is constant across time and space. See section B in the Appendix for a discussion.

country-level consumption and inequality.[6] Second, we employ this variable to train random forest models that use a wide variety of predictors – including nightlights, as well as a cell's remoteness, geography, disease environment, CO2 emissions, population, and characteristics of the ecosystem, among others – to predict consumption in the DHS cells and to assess prediction accuracy. Third, we use the trained models to create estimates of log consumption per capita across 10x10km cells. We then estimate poverty rates at the cell level by using a non-parametric approach that incorporates information from the nearly one million household surveys from DHS. We call the resulting estimates the Spatial Economic Development ("SED") data.[7].

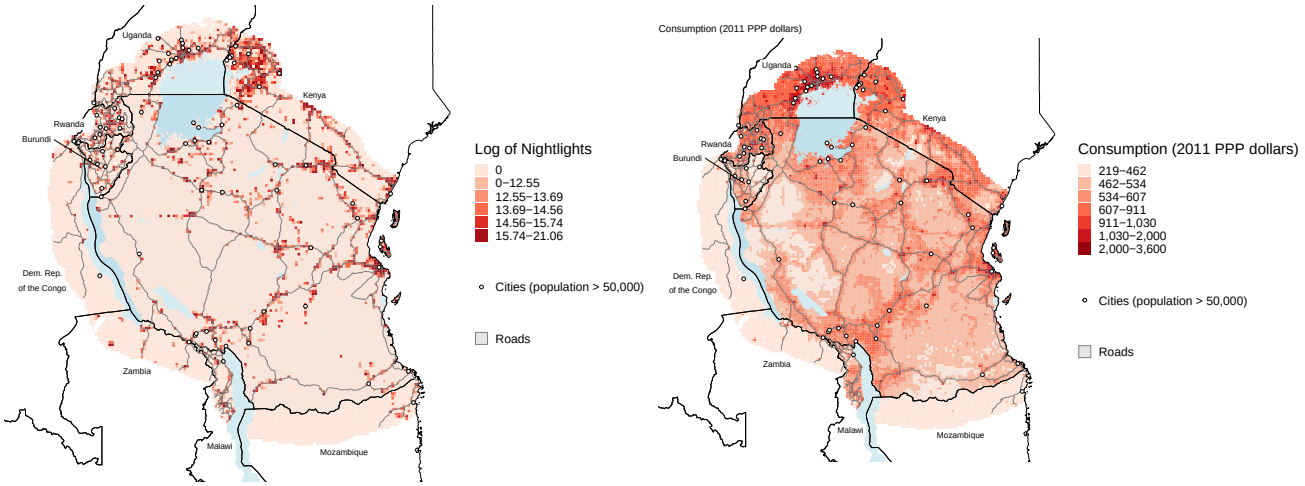
This procedure results in a novel dataset providing annual measures of per capita consumption and poverty rates at the cell level across 42 sub-Saharan African countries from 2003 to 2018. As we detail below, we improve on prediction accuracy with respect to previous papers, and do so at much lower computational cost. In fact, our entire analysis is executable in STATA, underscoring its feasibility for low-resource computing environments. Thus, the new data serves as a proof of concept and demonstrates the potential for our methods to be readily applied in generating new data for all developing countries.

Figure 1 illustrates how the SED data differ from nightlights. Panel (a) shows a map of nightlights in Tanzania (a country at about the 33rd percentile of development in Africa) in 2013 (a year when the more accurate VIIRS data is used to measure nightlights).[8] The most striking feature of the map is the vast swaths of darkness – 89% of cells in Tanzania are dark, and population estimates suggest 54% of Tanzanians live in these dark cells. The map also indicates that lit areas are almost exclusively located in larger cities and along the major roads, and that there is little differentiation in brightness levels when one crosses a national border. It is also important to note that though there is variation in light intensity across the 11% of cells that are lit, this variation is essentially uninterpretable given the metric used to measure

[6]See https://dhsprogram.com/Data/ and https://pip.worldbank.org/ for the DHS and the WB-PIP data, respectively.

[7]The new data can be visualized at https://www.spatial-economic-development.com/ where detailed consumption and poverty maps for the years 2003-2018 can be found

[8]The Visible Infrared Imaging Radiometer Suite (VIIRS) data, operational since 2012, provides more accurate and higher-resolution nightlight data compared to the older Defense Meteorological Satellite Program (DMSP) data. VIIRS offers improved sensitivity to low light levels and better calibration, making it more suitable for detailed monitoring and analysis of nighttime lights.



(a) Log of Nightlights

(b) Consumption per capita

Figure 1. Maps of nightlights and consumption in Tanzania. Panel (a) shows the map of (log) nightlights in 2013. The value 1 is added to each nightlight score so that when taking logs, dark areas on the map have a value of 0. Panel (b) shows the map of consumption per capita in 2013 using the SED estimates (RF-2) developed in this paper.

light intensity. Finally, the map reminds us that measures of nightlight intensity represent the same level of economic development no matter where the nightlight is measured. Thus, a cell with a nightlight measure of, say, 12.55 is assumed to have the same level of economic development whether this cell is in a city or the country, or whether the cell is in Tanzania, in a richer country (like Rwanda) or in a poorer one (like the DRC).

Panel (b) in the figure shows a map of consumption per capita using the new SED data. In addition to the fact that the cells are denominated in dollars, there are a number of clear differences from the nightlights map. These differences include that the consumption map (i) shows considerable variation within dark areas and within lit areas; (ii) reveals largely dark regions that have relatively high levels of development, such as southwest area of Tanzania near Lake Malawi; (iii) indicates how consumption changes as one moves along or away from highways, or away from cities; and (iv) shows that differences in consumption per capita are large across some national border areas but small across others.

We use the SED data to make our second contribution, which is to demonstrate that measurement error in nightlights is non-classical and that results from statistical models using nightlights as disaggregated measures of development therefore produce biased estimates,

regardless of whether nightlights are used to construct a dependent or an independent variable. To illustrate how non-classical measurement error in nightlights produces bias, we reconsider two prominent papers that use nightlights to study the effect of institutions on political development. Michalopoulos and Papaioannou (2013) use nightlights as a proxy for economic development to show that centralized ethnic institutions positively affect the latter. Michalopoulos and Papaioannou (2014) similarly use nightlights to show that strong national institutions (related to rule of law and corruption control) have no causal effect on national development. We use the data and statistical models from these two papers but substitute the SED measures of consumption for nightlights. Since predictions from machine learning models often suffer from nonclassical measurement error, we use a simple approach to eliminate such error from the SED estimates. We then use the corrected measure of consumption in the regressions. The results are the reverse of those in the original papers, with no effect of centralized ethnic institutions on development, and a strong positive effect of national institutions on development.

Why the change in results? We argue that non-classical measurement error (NCME) in nightlights plays a central role. More specifically, we argue that centralized ethnic institutions are positively correlated with the measurement error in nightlights, causing an upward bias in the coefficient estimates for centralized ethnic institutions. This explains how Michalopoulos and Papaioannou (2013) can obtain significant results when the true effect of ethnic institutions is likely zero. Similarly, we argue that national institutions are negatively correlated with nightlights measurement error, causing a downward bias in the coefficient estimates for the variables measuring these institutions. This explains how Michalopoulos and Papaioannou (2014) can obtain null results when in fact national institutions likely have a positive effect on development.

The problem of bias is not specific to these two papers. The NCME arises largely because so many cells in the developing world are dark, making the problem endemic to the use of nightlights as a spatially disaggregated measure of development. Without better data than nightlights, it is very difficult to know whether the bias resulting from the NCME is one of amplification or attenuation. Thus, there is a general problem inherent to making

inferences from models that use nightlights as either an outcome or explanatory variable. The dollar-denominated SED data, corrected to eliminate NCME, not only makes possible clear substantive interpretation of results, it also makes it possible to avoid the problem of bias inherent to nightlights.

The paper is organized as follows. Section 2 describes the problem of NCME in nightlights, including a discussion of the bias created when nightlights are used in regression models. Section 3 introduces the mathematical framework that we employ to construct the new training variable. It then describes the construction of this variable and presents evidence using both micro and macro data to validate the new measure. Section 4 describes the methods and the data used to estimate prediction models for the new training variable. It then assesses the prediction accuracy of the models. The section also describes our approach to estimating poverty rates in cells, and presents evidence that the cluster-level predictions, when aggregated to the national level, result in measures of consumption and poverty that are closely aligned with external measures of these national aggregates. Section 5 presents the SED data for 42 countries over time, and it describes evidence that these data are capturing within-country variation in a meaningful way.

Sections 6 and 7 focus on using the new data in regression analysis. Section 6 describes the problem of non-classical measurement error in the estimates of consumption, which can be seen in the negative correlation between the level of consumption in a cell and its associated predicted error. We then describe our approach for ridding the data of this error. Section 7 uses the SED data (with the NCME eliminated), to re-estimate models in Michalopoulos and Papaioannou (2013) and Michalopoulos and Papaioannou (2014). This exercise allows us (a) to illustrate the problem of biased estimates due to non-classical measurement error in nightlights, (b) to describe how SED data can be used to estimate the substantive size of the effects of institutions on economic outcomes, and (c) to conclude that strong national institutions have a very large effect on development whereas ethnic institutions do not. Section 8 concludes. Additional information is provided in the Appendix.

2. The problem of non-classical measurement error in nightlights

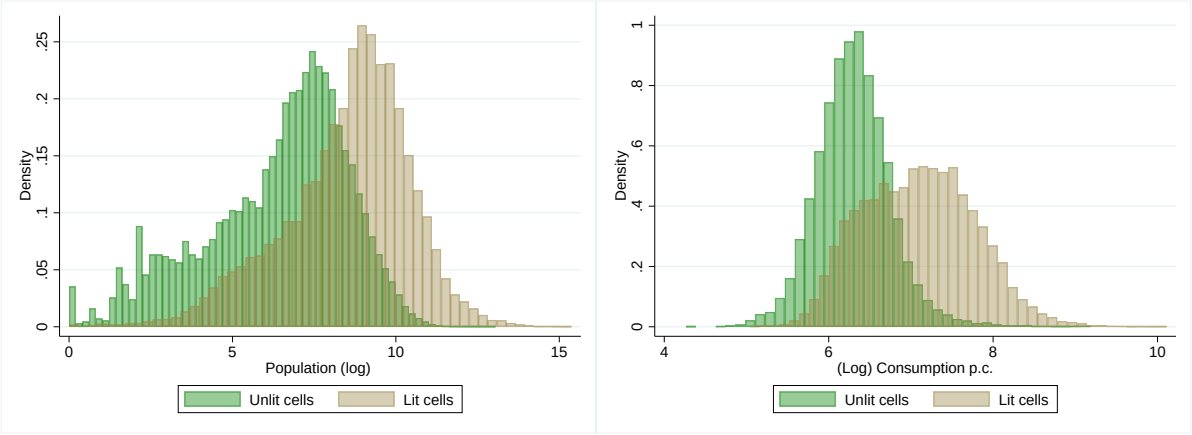
It is well-known that nightlights are measured with substantial error due to various factors, including changes in satellite technology, sensor saturation, overglow effects, and other issues (see Gibson et al. (2020) for a comprehensive review). Here we focus on the problem of non-classical measurement error in nightlights, which largely stems from the fact that for the vast majority of territory in the developing world, satellites detect no light at night. This problem was noted by Chen and Nordhaus (2011, p. 8594), who wrote that *"... luminosity data do not allow reliable estimates of low-output-density regions largely because the level of stable lights is too low to be distinguished from the background lights and is set at zero."* In sub-Saharan Africa, for instance, around 92% of the 10x10km cells are dark in the period 2006-2018.[9]

In what follows, we argue that this "problem of darkness" extends beyond mere data censoring; it involves significant misclassification, where economically prosperous areas are mistakenly identified as poor, and vice versa. This leads to nonclassical measurement error in NL data, which poses a considerable challenge for research employing NL in regression analysis. The remainder of this section discusses why the error in NL is nonclassical and summarizes the problems arising when using NL in standard statistical models.

Nonclassical measurement error in NL. Why should we expect NL to suffer non-classical measurement error? To address this question, it is useful to first recall the difference between the two types of measurement error in the NL context. Let y^* be the "*true*" measure of economic well-being, and define nightlights as a noisy proxy for y^*, with an additive error such that $NL = y^* + u$. Classical measurement error occurs when the true indicator y^* is uncorrelated with u. When y^* is correlated with u, the measurement error is nonclassical. This type of error is typically more difficult to deal with than is classical error, as we summarize below.

A central reason that NCME arises with nightlights is because the vast majority of cells in the developing world are dark. Importantly, the dark areas are full of people. Panel (a)

[9] There are two main sources of nighttime lights data. For the period before 2012, we utilize the Defense Meteorological Satellite Program (DMSP) dataset. In this dataset, 93% of the cells in sub-Saharan Africa are unlit, and this figure increases to 99% after correcting for the issue of overglow. For the period from 2013 to 2018, we employ the more accurate Visible Infrared Imaging Radiometer Suite (VIIRS) dataset. According to this dataset, 89% of the cells appear to be unlit for the years 2013-2018.



(a) Population in lit and unlit areas

(b) (Log) Consumption p.c. in lit and unlit areas

Figure 2. Relative frequency of (log) population (**Panel a**) and (log) consumption per capita (**Panel b**) in lit and unlit areas in Sub-Saharan Africa. To compute these graphs, Sub-Saharan Africa has been divided in cells of 10×10 km. The histogram in Panel (a) reflects the relative frequency distribution of population in lit versus unlit cells for the period 2006–2018. Panel (b) presents the histogram of this paper's measure of log consumption per capita in lit versus unlit areas for the subsample of cells for which geo-located survey data exist in this period (see section 3 for details). Yearly nightlights data comes from DSMP (blur data) for the period 2006-2012 and from VIIRS for the 2013-2018 period.

of Figure 2 shows the distribution of estimated cell population across lit and dark 10km^2 cells in 42 African countries from the period 2006-18 (which is the period for which we have DHS data). Although lit cells are obviously more populated, there is a large overlap in the two distributions. In 2018, existing population estimates suggest that 51.4% percent of the population lived in dark cells.

Since roughly half the population lives in dark areas, one should expect that there is substantial economic activity in these areas. This can be seen in panel (b) of the figure, which shows the distribution of (log) consumption per capita in dark and lit cells for the subset of country-years for which we have surveys to create training data.[10] Forty-five percent of the surveyed locations are "dark," and though lit pixels are richer on average, confirming the expected connection between nightlights and economic development, there is remarkable overlap of the two distributions and a wide range of consumption values within both sets of cells.

[10]The log-consumption data has been computed by combining information from 85 geo-located surveys from the DHS and country-level consumption per capita data from the World Bank; see section 3 for details.

The link between NCME in nightlights and the "problem of darkness" is easy to appreciate if one first considers the case where y^* is binary, i.e., y^* equals 1 when a latent variable of well-being exceeds a certain threshold. Consider also a binary version of NL, where all strictly positive NL values are set equal to 1. Since $NL = y^* + u$, u can only take three values in this simple example: $u = 0$ (if there is no misclassification), $u = -1$ (false negative case, where $y^* = 1$ and $NL = 0$), and $u = 1$ (false positive case, where $y^* = 0$ and $NL = 1$). As is well-known, misclassification in binary variables always results in nonclassical measurement error (Meyer and Mittag, 2017). For the case of nightlights just described, there is obviously a negative correlation between u and y^*.

Because of the vast areas of darkness, when we consider continuous values in NL and y^*, the reasoning from the binary case still operates. Given the problem of darkness, $NL = 0$ for more than 90% of the territory and, as is clear from Figure 2, there are many people and much economic activity in these dark spaces. This implies that $y^* = -u$ for the vast majority of cells, thus ensuring a negative correlation between y^* and u.

Using NL in Regression Analysis. The presence of non-classical measurement error has significant implications for the interpretation and validity of OLS and IV estimators (see Bound et al. (1994) for a comprehensive treatment). These implications affect the way we understand the results presented in section 7, and so we briefly review them here.

We first consider the case where NL is employed as a dependent variable in a spatially disaggregated data set. Suppose we would like to estimate the (true) model $y^* = \beta x + \epsilon$, but $NL = y^* + u$ is used in place of the unobserved y^*. We therefore estimate $NL = \beta x + (\epsilon + u)$. Assume for simplicity that $\beta \geq 0$ (otherwise, redefine x accordingly), that all variables are in deviations from their mean, and that x is exogenous. It follows that the bias δ of the OLS estimator is given by:

$$\delta = \hat{\beta} - \beta = \frac{cov(x'u)}{var(x)} + o_p(1). \tag{1}$$

In the classical case, u is assumed to be uncorrelated with y^* and x and therefore $\delta \xrightarrow{p} 0$. Thus, OLS coefficients are consistent, although they will be less precise due to increased variability.

In the non-classical case, the problems are considerable: since u is correlated with y^*, it is also likely to be correlated with x (as y^* and x are in principle related). This implies that $\delta \xrightarrow{p} 0$. Furthermore, the sign of the bias is determined by the sign of the correlation between x and u, which is often unknown. Since this simple example assumes that $\beta \geq 0$, a negative correlation will generate an *attenuation* bias whereas a positive correlation will generate an *amplification* bias.

Next consider the case where NL is a regressor. We wish to estimate the (true) model $z = \beta y^* + \varepsilon$ but instead estimate $z = \beta NL + (\varepsilon - \beta u)$, which implies

$$\hat{\beta} = \Big(1 - \frac{cov(y^*, u) + \sigma_u^2}{\sigma_{NL}^2}\Big)\beta + o_p(1) = (1 - \gamma)\beta + o_p(1). \tag{2}$$

In the classical error case, endogeneity issues arise and OLS coefficients are inconsistent. However, as $cov(y^*, u) = 0$, the coefficients are always biased towards zero (provided $\sigma_u^2 > 0$). Given the presence of this attenuation bias, researchers can interpret estimated coefficients as lower bounds on the true relationships between the variables.[11] But the problems are again more thorny when the error is non-classical. Assuming that $cov(y^*, u) < 0$, the bias is attenuating if $\gamma > 0$, which will occur if $var(u) > -cov(y^*, u)$. As $cov(y^*, u)$ becomes more negative, attenuation can become sufficiently severe that the sign of the estimated coefficient will be reversed. If $\gamma < 0$, (i.e., when $var(u) < -cov(y^*, u)$), the bias is amplifying.

Can this problem be solved using an instrument for the mismeasured regressor? This is an appropriate solution when the error is classical. Unfortunately, in the nonclassical case, finding suitable IVs is much more difficult. In this case, the endogeneity problem might not be confined to the mismeasured variable exclusively, as in the classical case. The reason is that the error term u is correlated with development, y^*, and since in multiple regression other regressors are typically also correlated with y^*, those regressors can also become endogenous. Consequently, using IV estimation to address the NCME when NL is a regressor might require identification of appropriate instruments for all the regressors in the model that are correlated with y^*. But not only are more IVs needed, finding valid IVs is more complicated in the

[11] Furthermore, when the measurement error is limited to a single regressor, it is possible to obtain an upper bound of the true coefficient by computing the reverse regression (i.e., using the mismeasured variable as a dependent variable). This classical finding, as documented by Frisch in 1934, provides valuable estimated bounds on the coefficients.

nonclassical case. This is due to the fact that a valid IV needs to be correlated with the endogenous variable(s) but *uncorrelated* with u. But since u and y^* are in fact correlated, the task of finding instruments that satisfy both conditions is formidable.

In summary, when conducting regressions using NL as either a dependent or independent variable, inconsistent coefficients are likely to emerge due to nonclassical measurement error. This inconsistency can result in attenuation or amplification bias, making the interpretation of estimates challenging. Furthermore, attempting to mitigate this issue through IV estimation is difficult, as nonclassical measurement error complicates considerably the search for valid instruments.

3. A new measure of spatial consumption per capita

This section presents a novel measure of per capita economic well-being, one that has high spatial resolution and that is denominated in an easily interpretable metric. The measure is created using supervised machine learning, which involves training a computational model on a dataset that includes both the geocoded target variable, known as the *training variable* and the geocoded predictors (also known as *features*). We use random forests, which learn by recognizing complex and potentially highly nonlinear relationships between the features and the training variable.

The primary challenge associated with using machine learning to this end is to identify an appropriate variable for training the model. The training variable must be geographically fine-grained, have broad temporal and spatial scope to ensure comprehensive learning by the model, and be denominated in a metric useful for economic research, such as consumption or income. Unfortunately, in the context of sub-Saharan Africa and many other developing parts of the world, this appropriate training variable does not exist. Several data sources are available, but none that combine adequate geographic and temporal scope with a measure of economic well-being denominated in a useful metric. For instance, the Living Standards Measurement Study (LSMS) uses surveys to measure household consumption and asset ownership, but in sub-Saharan Africa the number of usable surveys containing geocoded enumeration areas with sufficient respondents in each is too small to adequately train a

predictive model. The Demographic and Health Surveys (DHS) offer broad temporal and spatial scope and have geocoded enumeration areas with sufficient respondents. But the only measure of economic well-being that can be derived from these surveys is an index of asset ownership, which is expressed in arbitrary units. Finally, the World Bank (WB-PIP) has amassed an extensive collection of surveys on income and/or consumption levels from developing countries, but the publicly accessible data are mostly limited to different aspects of the country-level distribution, such as its mean, the Gini coefficient, poverty rates or deciles.[12]

Our solution to this problem is to develop a framework for translating an asset index into a measure of consumption. This allows us to leverage the broad temporal and spatial scope of the DHS surveys, but to do so in a way that utilizes a training variable that contains interpretable economic information. In the remainder of this section, we first introduce a simple mathematical framework designed to enable the computation of a novel measure of consumption by integrating existing datasets. In so doing, we make clear the assumptions underlying the construction of the new measure, and discuss the consequences when the assumptions are violated. We then describe how we implement this framework and conclude the section by describing evidence to validate the measure, including a critical examination of the key assumptions within the framework.

3.1. Deriving the Training Variable: Mathematical Framework. We assume the existence of a latent variable, y^*, which represents the "true" level of economic well-being. While we interpret y^* as the log of per capita consumption (expressed in log dollars for the sake of this discussion), it could alternatively embody other economic metrics such as income or expenditure.[13]

While y^* is unobserved, two proxies for this quantity exist and can be (at least partially) observed: y^C and y^A. y^C is a consumption index, also measured in log dollars. The second indicator, y^A, is an index of asset ownership.[14] We adopt a framework similar to Chen and

[12]Recently, the World Bank has begun to grant access to microdata via the Poverty and Inequality Platform. However, the number of surveys for which these data are accessible remains quite limited. For further information and data availability, see at https://pip.worldbank.org/home.

[13]Throughout this paper, unless explicitly indicated otherwise, all variables are expressed in logarithmic form.

[14]In practice, the variable y^A is observed at the individual level across numerous country years, utilizing DHS surveys. However, for y^C, only measures of centrality and dispersion are available at the country-year level, sourced from the WB-PIP data.

Nordhaus (2011) and Pinkovskiy and Sala-i-Martin (2016) and assume that both y^A and y^C are linearly related to y^* in the following way.[15]

Assumption A. The variables y^C_{ict} and y^A_{ict} are related to y^*_{ict} as follows:

$$y^C_{ict} = y^*_{ict} + \epsilon^C_{ict}, \tag{3}$$

$$y^A_{ict} = \alpha_{ct} + \beta_{ct} y^*_{ict} + \epsilon^A_{ict}, \quad \beta_{ct} > 0, \tag{4}$$

where i indexes individuals, c indexes countries, and t indexes time. The errors ϵ^C_{ict} and ϵ^A_{ict} have zero mean, are mutually uncorrelated and are uncorrelated with y^*_{ict}.

Equation (3) expresses the assumed relationship between existing consumption data and y^*. It implies that although y^C_{ict} is measured with error, it is unbiased; i.e., $E_{ct}(y^C_{ct}) = \mu^C_{ct} = E_{ct}(y^*_{ct}) = \mu^*_{ct}$, where $E_{ct}(.)$ represents the expected value over the distribution of individuals for country c at time t. Equation (4) expresses the assumed relationship between an asset index, y^A, and y^*. y^A_{ict} is measured in different units than y^C_{ict} and y^*_{ict} and, since these units are arbitrary, we assume without loss of generality that y^A_{ct} has a mean of zero and a standard deviation equal to 1 for each country c and year t (otherwise, redefine α_{ct} and β_{ct} accordingly).

Using equation (4), define a new proxy for y^* as $\widetilde{y^*_{ict}} = (y^A_{ict} - \alpha_{ct})/\beta_{ct}$, which can be written as

$$\widetilde{y^*_{ict}} = y^*_{ict} + \widetilde{\epsilon_{ict}}, \text{ where } \widetilde{\epsilon_{ict}} = \epsilon^A_{ict}/\beta_{ct}. \tag{5}$$

The proxy variable $\widetilde{y^*_{ict}}$ has two key properties. First, it is an unbiased proxy for y^*_{ict} that is measured in the same units as y^*_{ict} (i.e., log dollars). Second, if α_{ct} and β_{ct} were known, $\widetilde{y^*_{ict}}$ could be observed at the micro level provided y^A_{ict} is observable too. To identify the parameters α_{ct} and β_{ct}, note that combining equations (3) and (4) yields

$$y^C_{ict} = \frac{(y^A_{ict} - \alpha_{ct})}{\beta_{ct}} + (\epsilon^C_{ict} - \epsilon^A_{ict}/\beta_{ct}). \tag{6}$$

[15] See also Hruschka et al. (2015) for an alternative strategy to transform asset-based indices. Their approach is developed under very stringent distributional assumptions since it adopts parametric assumptions on the distribution of the asset-based index, as opposed to the approach presented in this section.

Given that $E(y^A_{ict}) = 0$ and $Var(y^A_{ict}) = \sigma^2_{y^A_{ct}} = 1$, equation (6) implies that

$$E(y^C_{ct}) = \mu_{y^C_{ct}} = \frac{-\alpha_{ct}}{\beta_{ct}}, \quad \text{and} \tag{7}$$

$$Var(y^C_{ct}) = \sigma^2_{y^C_{ct}} = 1/\beta^2_{ct} + (\sigma^2_{\epsilon^C_{ct}} - \sigma^2_{\epsilon^A_{ct}}/\beta^2_{ct}), \tag{8}$$

where $\sigma^2_{\epsilon^i_{ct}}$ denotes the variance of ϵ^i_{ct} for $i = \{C, A\}$.

To simplify the expressions of α_{ct} and β_{ct}, Assumption B is very useful (though we examine the consequences of its violation below).

Assumption B. The variances of the measurement error in y^C_{ct} (the consumption variable) and in $\widetilde{y^*_{ct}}$ (the asset variable transformed by α_{ct} and β_{ct} into a measure of consumption) are similar in magnitude; i.e., $\sigma^2_{\epsilon^C_{ct}} - \sigma^2_{\epsilon^A_{ct}}/\beta^2_{ct} \approx 0$.

Under Assumption B, equation (8) simplifies to $Var(y^C_{ct}) = \sigma^2_{y^C_{ct}} = 1/\beta^2_{ct}$. Therefore, this assumption makes it possible to obtain expressions for α_{ct} and β_{ct} that depend only on *country level* moments of y^C_{ct}:

$$\beta_{ct} = 1/\sigma_{y^C_{ct}}, \quad \text{and} \quad \alpha_{ct} = -\beta_{ct}\mu_{y^C_{ct}} \Rightarrow \alpha_{ct} = -\frac{\mu_{y^C_{ct}}}{\sigma_{y^C_{ct}}}. \tag{9}$$

Equation (9) implies that consistent estimates for α_{ct} and β_{ct} can be computed by plugging consistent estimates of $\mu_{y^C_{ct}}$ and $\sigma_{y^C_{ct}}$ in those expressions. An estimate for $\widetilde{y^*_{ict}}$, which we will call $\widehat{y^*_{ict}}$, is therefore obtained by replacing α_{ct} and β_{ct} with their corresponding estimates:

$$\widehat{y^*_{ict}} = \frac{(y^A_{ict} - \hat{\alpha}_{ct})}{\hat{\beta}_{ct}} = \widehat{\sigma_{y^C_{ct}}} y^A_{ict} + \widehat{\mu_{y^C_{ct}}}. \tag{10}$$

To summarize, assumptions A and B make it possible to construct a new proxy of (log) consumption at the individual level, $\widehat{y^*_{ict}}$, by combining two types of (existing) datasets, one on individual-level asset ownership (DHS) and another that provides the first and second order moments of the country-level consumption distribution (WB-PIP).

3.2. Violations of key assumptions. We now consider the consequences of violating the two key assumptions in the mathematical framework.

Violation of Assumption A. Assumption A has two key elements. The first is that through their connection with y^*, y^C_{ict} and y^A_{ict} are linearly related (eq. 6). The second is that y^C_{ict} is an unbiased proxy, i.e., $E_{ct}(y^C_{ct}) = \mu^*_{ct}$.

If Assumption A does not hold, there is no reason to expect that y^C_{ict} and y^A_{ict} should have an empirical relationship that is linear. We will examine this implication in section 3.4 using data from the LSMS, as these surveys contain information about y^C_{ict} and y^A_{ict} for the same household.

If y^C_{ict} is not an unbiased proxy of y^*_{ict}, $\widetilde{y^*_{ict}}$ will not be unbiased either. More specifically, assume that $y^C_{ict} = \delta_0 + \delta_1 y^*_{ict} + \epsilon^C_{ict}$ (and, therefore, $E_{ct}(y^C_{ct}) = \delta_0 + \delta_1 \mu^*_{ct}$), which implies that $E_{ct}(\widetilde{y^*_{ct}}) = \delta_0 + \delta_1 \mu^*_{ct}$.[16] It follows the bias of $\widetilde{y^*_{ict}}$ will be similar to the bias of y^C_{ict}, but no worse. Unfortunately, since data on y^* is not available, it is not possible to test this assumption. This fact highlights the importance of using good quality country-level data to transform the variables, as the "transformed" consumption index inherits the biases of the consumption data used to compute it.

Violation of Assumption B. If Assumption B does not hold (that is, $\sigma^2_{\epsilon^C_{ct}} - \sigma^2_{\epsilon^A_{ct}}/\beta^2_{ct} \not\approx 0$), then $\bar{\beta}_{ct}$, defined as $\bar{\beta}_{ct} = 1/\sigma_{y^C_{ct}}$, is no longer equal to β_{ct}. Define $\bar{y}^*_{ict}$ as the proxy obtained by using $\bar{\beta}_{ct}$ instead of the true β_{ct}. By equation (7), it is computed as $\bar{y}^*_{ict} = y^A_{ct}/\bar{\beta}_{ct} + \mu^C_{ct}$.

The violation of Assumption B does not affect the unbiasedness of the new proxy; i.e., $E_{ct}(\bar{y}^*_{ct}) = \mu^C_{ct} = \mu^*_{ct}$ since $E_{ct}(y^A_{ct}) = 0$. The proxy $\bar{y}^*_{ict}$ is related to y^*_{ict} through the following equation:

$$\bar{y}^*_{ict} = \sigma^C_{ct} y^A_{ict} + \mu^C_{ct} = (\alpha_{ct}\sigma^C_{ct} + \mu^*_{ct}) + \sigma^C_{ct}\beta_{ct} y^*_{ict} + \epsilon^A_{ict}\sigma^C_{ct}. \tag{11}$$

Violation of Assumption B therefore implies that the intercept in equation (11), $\alpha_{ct}\sigma^C_{ct} + \mu^C_{ct}$, is different from zero, and the slope, $\sigma^C_{ct}\beta_{ct}$, is different from 1. This implies that the validity of Assumption B could be tested by regressing the transformed asset proxy on y^C_{ict} and then examining whether the intercept is zero and the slope is one.[17] We conduct this test in section 3.4.

[16] To see this, recall that $\widetilde{y^*_{ict}} = \frac{y^A_{ict}}{\beta_c} - \frac{\alpha_{ct}}{\beta_{ct}}$. Since $E_{ct}(y^A_{ct}) = 0$, it follows that $E_{ct}(\widetilde{y^*_{ct}}) = 0 - \frac{\alpha_{ct}}{\beta_{ct}} = \mu^C_{ct} = \delta_0 + \delta_1 \mu^*_{ct}$.

[17] Notice that under Assumption A, $y^C_{ict} = y^* + \epsilon^C_{ct}$ and therefore, one can use y^C_{ict} in place of y^*_{ict} in equation (11).

3.3. Constructing the training variable: Cluster-level measures of consumption. This section briefly describes the steps involved in constructing the consumption measure used as a training variable. A more detailed explanation can be found in Appendix B.

The first step is to construct the individual-level indicator, $\widehat{y^*_{ict}}$. We use 85 DHS surveys from 29 sub-Saharan African countries during the period 2006-2018. The surveys comprise over 900,000 respondents who are heads of households, and for each we calculate $\widehat{y^*_{ict}}$ as defined in eq. (10).[18] This requires an asset based index, y^A_{ict}, as well as country-year estimates of μ^C_{ct} and σ^C_{ct}. For each survey, we estimate a principal components model using all households and a variety of asset variables that are present in the given survey (see Appendix B.1 for a description of the assets). The log of the index from the principal components model yields y^A_{ict}.

The second step is to use the World Bank's WB-PIP data on average consumption and the Gini coefficient for the country-year of the survey to obtain estimates of μ^C_{ct} and σ^C_{ct}, the mean and the standard deviation of *log*-consumption.[19] Importantly, the WB-PIP data are based on surveys, and thus avoid biases that can occur in national accounts data (see, e.g., Martinez 2022). We use the estimates of y^A_{ict}, μ^C_{ct} and σ^C_{ct} to compute $\widehat{y^*_{ict}}$ as described in section 3.1, eq. (10). The resulting measure, $\widehat{y^*_{ict}}$, is expressed as the log of consumption in 2011 PPP dollars.

DHS groups respondents into geocoded clusters, the DHS term for an enumeration area, which we index by r. An advantage of the DHS survey methodology for our purposes is that the surveys include a sufficient number of households in each cluster to obtain meaningful cluster averages.[20] Thus, the third step is to compute cluster-level averages of $\widehat{y^*_{ict}}$, yielding $\widehat{y^*_{rct}}$, our training variable, expressed as the log of consumption per capita in 2011 PPP dollars. The 85 surveys in our data contain 34,483 clusters that we can use in prediction. Since DHS households are within 5km of the cluster centroid (which is known), we assign each cluster to a 10x10 square kilometer cell that has as its centroid the latitude and longitude of the cluster.

[18] DHS data are collected at the household level. To obtain per capita measures, we apply standard transformations; see Appendix B.1 for details.

[19] Appendix B.2 describes how these estimates are obtained.

[20] We include only clusters that have at least 16 households. The mean cluster size is 26 households and the median is 25.

3.4. Validating the training variable. This section assesses the validity of the framework described in section 3.1 for translating asset indices into measures of consumption. We begin by considering the direct empirical implications of violating Assumptions A and B, as discussed in section 3.2. The most direct way to assess these implications is to utilize surveys that have measures of both consumption and assets. The DHS surveys we use as the main source of asset indices in this paper have no measure of consumption. Thus, we begin by using the Living Standards Measurement Study (LSMS).[21]

Validation using LSMS data. We focus on sub-Saharan countries in our sample time period (2006 onwards) that contain measures of respondents' consumption and asset ownership. We use one survey per country, and to homogenize comparisons, when more than one survey exists in a country, we select the survey closest to the years 2012-13. The result is a dataset composed of 49,062 household from seven countries.[22] For each household, we compute a (log) consumption index and a (log) asset-based index, transforming the asset index as described in section 3. We use the LSMS survey log-consumption mean to measure $\mu_{y_{ct}^C}$ and we use the standard deviation of this variable to measure $\sigma_{y_{ct}^C}$.

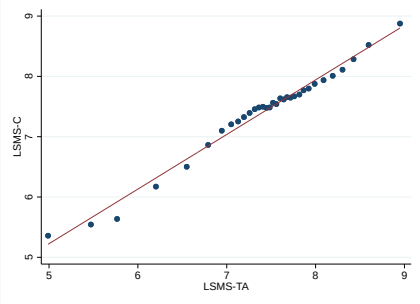
Given our focus on cluster-level data in the prediction exercises below, we focus on "cluster" level LSMS data, which is created using LSMS enumeration areas. In each enumeration area, we average the respondents' consumption scores and translated asset scores to create the cluster-level measures.[23] Let LSMS-C be the measure of consumption in an LSMS cluster, and let LSMS-TA be the measure of consumption using the transformed asset variable. There is a total of 2,461 clusters in the seven countries.

As underlined in section 3.2, Assumption A implies *there should be a* ***linear*** *relationship between* y_{ct}^C *and* $\widehat{y_{ct}^*}$. Panel (a) in Figure 3 shows the binned scatter plot of LSMS-C and LSMS-TA. Though there are deviations from linearity, these deviations are quite small, suggesting that the relationship between the two variables is in fact quite close to linear.

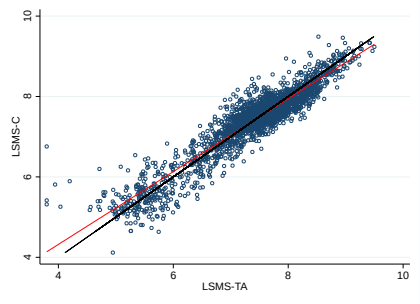
[21] See https://www.worldbank.org/en/programs/lsms. Note the LSMS surveys cannot be used as training data due to the paucity of surveys and to the frequent absence of geocoded enumeration areas of the appropriate size.

[22] The seven countries with the required measures are Burkina Faso (2014), Ghana (2010), Malawi (2013), Niger (2011), Nigeria (2018), Tanzania (2012) and Uganda (2013).

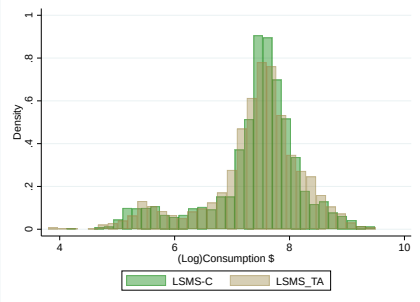
[23] Due to data limitations with LSMS enumeration areas, we impose a minimum of 10 households in a cluster for inclusion of the cluster.



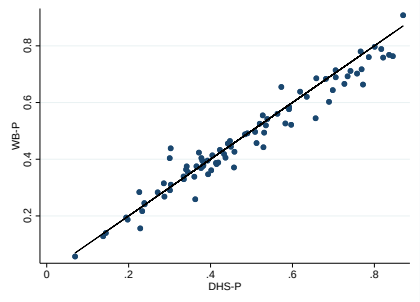
(a) Binned scatter plot of LSMS-C vs. LSMS-TA (both measured as log consumption dollars per capita)



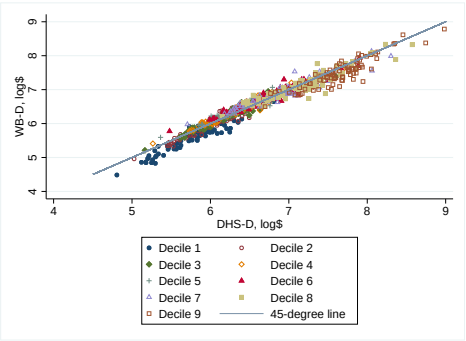
(b) Scatter plot of LSMS-C vs. LSMS-TA (both measured as log consumption dollars per capita), along with 45-degree line (in black) and regression line (in red)



(c) Histogram of LSMS-C and LSMS-TA, (both measured as log consumption dollars per capita)



(d) Scatter plot of country-level poverty rates using WB-PIP (WB-D) and DHS transformed assets (DHS-D) , along with 45-degree line. Poverty line is $1.90 (2011 PPP dollars).



(e) Scatter plot of mean of decile values using WB-PIP (WB-D) and DHS transformed asset-based (DHS-D) (both measured in log consumption dollars per capita), along with 45-degree line

Figure 3. Validating the transformation of asset indices into measures of consumption.

Assumption B implies that *if we regress the transformed asset proxy on y^C_{ict}, the intercept should be near zero and the coefficient for y^C_{ict} should be near one*. Panel (b) in Figure 3 shows the scatter

plot of LSMS-C against LSMS-TA, along with the red regression line and the black 45-degree line. Although the red line has a slightly flatter slope than the 45-degree line (suggesting the transformed asset variable is compressing the tails of the distribution a bit), the two variables are very closely related. In the regression, the constant is 0.70 (p=.06), the slope is 0.91 (p=.000) and the adjusted R-squared is 0.87. Thus, these data suggest that Assumption B is not perfectly satisfied, but the departures from what the assumption requires are modest. Indeed, if the six outlying clusters (out of 2,461) from Ghana in the southwest corner of the scatter plot are excluded, the constant is no longer statistically significant (p=.18) and the slope coefficient is 0.92, with a 95% confidence interval that includes 1 (it ranges from .81 to 1.04), making it impossible to reject the hypothesis that the slope equals 1 (p=0.15).

It is also useful to consider the overall distributions of the two variables. If both assumptions are satisfied, and if the measurement error is not too large, then we should also expect the distribution of LSMS-C to be similar to the distribution of LSMS-TA. Panel (c) in the figure show the histograms of the two variables, and it reveals that the distributions are very similar to each other.

Validation using WB-PIP data. Next consider evidence regarding the validity of the mathematical framework obtained by comparing the distributions of the consumption variable provided by the WB-PIP and the new consumption variable calculated using DHS surveys. As noted in connection with panel (c), if assumptions A and B are satisfied and if measurement error is not too large, the distribution of consumption using a transformed asset variable should be similar to the distribution of consumption itself.

In addition to the mean of consumption and the Gini coefficient, both of which are employed to compute the new consumption index, WB-PIP provides additional information on the country-level distribution of individual consumption: the poverty rates and the values of each decile. Denote the poverty rate published by WB-PIP as WB-P. To calculate a country poverty rate from the new consumption proxy, which we denote DHS-P, we calculate for each survey the proportion of respondents with asset-based consumption scores that are below the poverty line of $1.90 per day. Panel (d) in Figure 3 presents the results. The poverty rates derived from the new consumption proxy are indeed very similar to the poverty rates

published by WB-PIP – the correlation is 0.98 – and this is true across the range of poverty rates that exist in these country-years, though there are modestly higher poverty rates using DHS-P in the higher ranges of poverty.

WB-PIP also publishes the consumption values at different deciles in the consumption distribution. As in the calculation of poverty rates, we can use the new consumption data to calculate the value of consumption at each decile in a survey. Let WB-D refer to decile values using WB-PIP and let DHS-D refer to decile values derived from the new data. Panel (e) of Figure 3 presents for each decile the scatterplot of the 85 values using WB-D against the values from DHS-D. The solid circles, for example, plot the values for the first decile, and there is a tendency for WB-D to have slightly lower values than DHS-D values in this decile. But the figure shows that there is a very strong relationship between the decile values from WB-D and the decile values derived from the transformed DHS asset variable. Indeed, the correlation between the two variables is at least 0.95 for 7 of the 9 deciles (and is 0.90 in decile 9 and 0.93 in decile 8). These strong relationships across all 9 deciles provide additional evidence that we should have confidence that transformed assets can be used to create an accurate distribution of consumption.

4. Prediction

This section lists the predictors (features) employed to estimate the prediction models (section 4.1), describes the algorithm and models employed (section 4.2), and assesses prediction accuracy (section 4.3). In addition, we aggregate cluster data on consumption and poverty (which we estimate at the cluster level using a non-parametric method described below) to the national level and compare these consumption and poverty aggregates with data on consumption and poverty from the World Bank (section 4.4).

4.1. Features. We use the following predictor variables, defined over 10 km^2 cells. Time-varying predictors are in bold. For a detailed description and data sources, please see Appendix C.

(1) **Nightlights.** Cell-level intensity of lights at night. Nightlights are measured using DMSP data for the period 2006-2012 and using VIIRS for the remaining years. We also

include a standardized version of three-year averages that comprises the entire time period, as described in Yeh et al (2020).

(2) *Core variables*. Cell-level variables describing different features of the cell that are related to:

   (a) *Geography.* Ecosystem type, ruggedness of terrain, elevation, latitude and longitude.

   (b) *Distances.* Distance from the pixel to the capital, a highway, the coast, a harbor, a protected area, a river, and catholic and/or protestant missions;

   (c) *Climatic variables/disease environment*: **Temperature, rainfall and malaria incidence**.

   (d) *Other indicators of economic activity*: **Population density** and **CO2 production**.

4.2. Algorithm: Random Forests. We use a random forest (RF) algorithm to predict the measure of log consumption in each cell. The RF algorithm is an ensemble method, i.e., it is made up of a large number of individual decision trees, each producing their own predictions. The random forest algorithm combines these individual predictions to produce a more accurate one. This is important because standard decision tree algorithms have the disadvantage that they are prone to overfitting. The ensemble design allows the random forest to avoid this problem.[24]

The RF algorithm has several advantages: it has impressive prediction accuracy; it performs well when using a relatively large number of predictor variables (as opposed to other methods, such as K-Nearest Neighbor); and it is much less computationally intensive than other approaches, such as neural networks. This ability to achieve accurate predictions at low computational cost is important. It allows scalability to a large number of country-years, and makes it possible to assess the robustness of estimates that emerge from any particular model. Moreover, these methods can be executed in popular statistical software like STATA, thereby facilitating their adoption by the broader research community with ease.[25]

We consider three RF models that differ in the predictors employed. Model RF-1 includes only nightlights. RF-2, our core model for generating data on cell consumption, includes

[24] Breiman (2001) provides a general description of random forests.

[25] See Scholao and Zou (2020) for details on the algorithm we employ in Stata to compute the predictions.

nightlights and the core variables. Finally, to gain additional insight into the role of nightlights versus the additional variables in the model, RF-3 removes nightlights from RF-2. To provide a comparison with previous literature, we also estimate models that include only nightlights as predictors using OLS and using the K-Nearest Neighbor algorithm.

Using RF requires the tuning of various hyperparameters to optimize performance. These hyperparameters, which include the number of trees in the forest, the depth of each tree, and the minimum number of samples required to split a node, among others, play a crucial role in the model's ability to learn from data without overfitting. See section D.1 of the Appendix for details regarding parameter tuning, as well as the hyperparameter values selected for each model.

4.3. Out-of-Sample Predictive performance. All evaluation is done on held-out locations. Specifically, for each survey s, a random forest model using clusters from all surveys other than s is estimated, and predictions are then obtained for the clusters in survey s. This approach replicates the real-world setting of making predictions where ground data do not exist.

To assess prediction accuracy, we focus primarily on mean square error (MSE) computed from the out-of-sample forecasts described above. To facilitate comparison with previous work (Yeh et al. 2020, Jean et al. 2016), we also display the R^2 of the out-of-sample predictions, which is computed as the square of the within-survey correlation between the training variable and the (out-of-sample) predictions. For each omitted survey s, we produce out-of-sample forecasts and compute the MSE and the R^2.

Table 1 reports the median values of the out-of-sample survey-level MSE and R^2 across the 85 out-of-sample predictions for the three models described above. The first three models use a random forest algorithm and different combinations of predictors, as described above. Three key findings from the table merit attention. First, prediction accuracy is high. The full model, RF-2, exhibits a MSE of 0.135 and an R^2 value nearing 0.72. Not only does this performance surpass existing benchmarks (e.g. Yeh et al. 2020), it also achieves this performance at a lower computational cost.[26] This efficiency suggests that the model's predictions are highly

[26] Appendix D.5 provides a direct comparison of model performance using our approach and that of Yeh et al. (2020), using the Yeh et al. training variable in both cases. We show that the RF approach with the rich set of

scalable.[27] Second, adding the core variables to nightlights substantially improves predictive performance. One can see this by comparing RF-1 (which uses only NL) with RF-2 (which adds all the "core" variables). The median MSE decreases by 32% when these additional predictors are added. Finally, the core variables are more informative than NL. This can be seen by comparing the performance of RF-3 (which excludes nightlights), with RF-1, which contains only nightlights.

One might hope that the poor results for RF-1 would improve by using a K-Nearest Neighbor (KNN) algorithm, which though computationally intensive, often performs well when there are a small number of predictors. Yeh et al. (2020), for example, do not use RF, but they find in a setting similar to ours that the KNN algorithm using only nightlights performs basically identically to a convolutional neural network (CNN) with only nightlights, or to a CNN using nightlights and daylight imagery.[28] The results from Model 4 show that the MSE associated with KNN is around 7% larger than that of RF-1, and much larger than that of RF-2. Finally, Model 5 aims to reflect the "naive" approach of using a linear function of NL as a predictor of economic activity. This model displays the worst performance, with a MSE that is about 2.5 times larger than that of the core model (RF-2).

	Median MSE	Median R^2
RF-1: NL	0.199	0.650
RF-2: NL, Core	0.135	0.716
RF-3: Core	0.141	0.680
Model 4: KNN with NL	0.242	0.579
Model 5: OLS with NL	0.323	0.391

Table 1. Prediction Accuracy. This table displays the median MSE and R^2 corresponding to the 85 sets of out-of-sample predictions estimated in five different ways. The first three models are estimated using a random forest algorithm and differ in the predictors included in the model. Core contains the core variables described in section 4.1. Models 4 and 5 include only NL as predictors and are estimated by KNN and OLS, respectively.

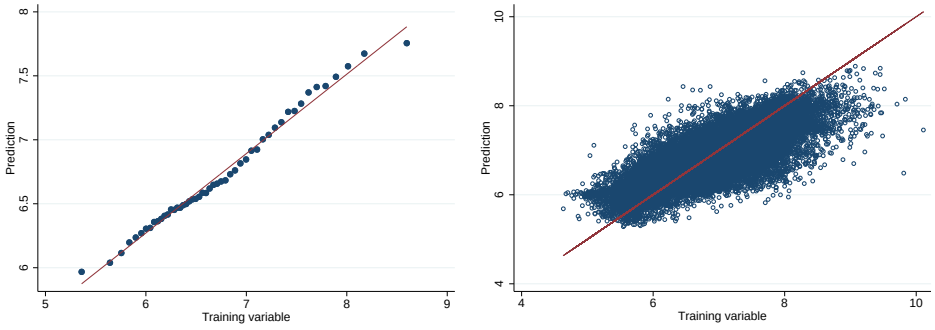
Figure 4 plots the training data against the predicted values from the core model, RF-2. The binned scatter plot in panel (a) shows that there is a tight linear relationship between

predictors provides more accurate estimates than the Yeh et al approach based on a much more computationally expensive convolutional neural network.

[27] For instance, Yeh et al. (2020) rely on computationally intensive neural networks. Coupled with their use of daylight imagery as a predictor, which is challenging to process, the scalability of their methodology is substantially limited.

[28] Model 4 in Table 1 uses the same nightlights predictors as RF-1 but estimates the model using KNN.

the training data and the prediction. Panel (b) shows that there is a tendency to slightly overestimate (log) consumption of the very poor and underestimate it for the very rich. The scatter plot therefore suggests that the RF estimates suffer non-classical measurement error (with the error negatively correlated with consumption). In section 6 below, we discuss this issue and provide a method for eliminating such error.



(a) RF-2: Binned scatter plot **(b)** RF-2: Scatter plot

Figure 4. Predicted versus Training values, Model RF-2. Panel (a) displays a binned scatterplot of predicted versus training data while Panel (b) presents the scatter plot containing all the data points.

Appendix D provides additional information about (a) the importance of different predictors in obtaining the estimates of log consumption (Appendix D.2) (b) prediction performance at the survey level (Appendix D.3); (c) the poor performance of the models that use only nightlights (Appendix D.4); and (d) a direct comparison of model performance using our approach with performance in Yeh et al. (2020) (Appendix D.5).

4.4. Predictive performance at the country level: national consumption and poverty. A second way to evaluate predictive performance is to aggregate the predictions of the cluster level data into national aggregates and to compare these aggregates with the WB-PIP national aggregates. Here we do this for consumption and poverty using our core model, RF-2. It is straightforward to compute the national consumption per capita by aggregating the cluster-level predictions of consumption and to compare the results with measures published by WB-PIP. In addition, the geographic distribution of poverty is of central substantive interest in developing parts of the world. But since $\widehat{y^*_{rct}}$ captures *average* consumption in the cluster r, it provides no information about the *within-cluster* distribution of consumption, and thus

provides no information about the cell's poverty rate. We therefore use a non-parametric approach for estimating the poverty rate for each cluster, based on the estimated consumption for each cluster. We then use the cluster poverty rates to estimate the national poverty rate, which can be compared with WB-PIP poverty rates.

Estimating cell poverty rates. One possible "naive approach" to estimating cell poverty rates is to assume that all individuals in a cell have the same level of consumption, so that all within-cluster distributions have a variance of zero. Under this assumption, the cell poverty rate would be 0 if the cluster's mean consumption were below the poverty line and would be 1 if the mean were above the poverty line. At the opposite extreme, one could assume that each cluster has its own distribution, with no clear relationship between mean consumption in a cell and its poverty rate. This assumption, however, would make it impossible to use the only available information – the estimate of cell consumption – to estimate cell poverty rates.

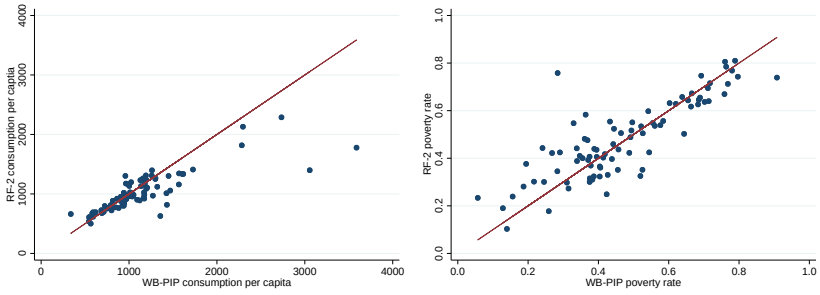
We adopt an intermediate position. The main assumption in our approach to estimating cell poverty rates is that all clusters with a very similar level of consumption per capita share the same distribution of consumption, and therefore share the same poverty rate. This assumption allows us to use household-level consumption information to compute within-cluster poverty rates for each of the groups.

The approach works as follows.[29] First, 100 groups of clusters are defined by the percentiles of the cluster-level distribution of consumption per capita. The first group, for example, includes all clusters that are in the first percentile of the distribution of consumption, obtained using all clusters in the training data. Second, each of the roughly 920,000 respondents in the DHS surveys are assigned to the group associated with the respondent's cluster mean income. Third, we use the individual consumption levels of the roughly 9,000 respondents in each group to compute the poverty rate of the group (by simply calculating the percent of the roughly 9,000 respondents in poverty, using the 2011 PPP poverty line of $1.90 per day). Finally, to calculate the poverty rate for a cluster, we assign the poverty rate for the cluster's group.

[29] Additional details regarding the estimation of cell poverty rates are provided in Appendix E.

National consumption per capita and poverty rates using WB-PIP and RF predictions. Figure 5 plots the national measures of consumption and poverty from WB-PIP against the measures obtained by aggregating the out-of-sample predictions from the the core prediction model, RF-2. Panel (a) depicts country-level log consumption per capita from WB-PIP and from RF-2, along with the 45-degree line. The estimated consumption tends to be higher using WB-PIP, especially in the richest countries. But the correlation between the two variables is 0.82, and though there are several exceptions, the preponderance of points on the graph are quite close to the 45-degree line. Thus, the predictions from RF-2 generate national estimates of consumption that are closely aligned with those from WB-PIP.

Panel (b) plots the WB-PIP poverty rates against the poverty rates derived from applying the non-parametric approach to the estimates from RF-2. The poverty rates from WB-PIP tend to be lower, but again there is a very strong relationship across the range of poverty rates, and the correlation of the two variables is 0.84. These macro-comparisons provide evidence that the accurate cluster-level predictions can be aggregated to provide sensible estimates of macro consumption and poverty, albeit estimates that are slightly more pessimistic about the level of development than those published by WB-PIP.



(a) Consumption p.c RF-2 (y-axis) vs. WB-PIP (x-axis)

(b) Poverty rate RF-2 (y-axis) vs. WB-PIP (x-axis) (using $1.90 poverty line and 2011 PPP dollars)

Figure 5. Country-level comparison. These figures display country-level consumption per capita and poverty rates from WB-PIP and out-of-sample predictions from RF the 85 country-years in our sample.

5. Spatial Economic Development ("SED") data for sub-Saharan Africa

The last step in data creation is to develop estimates for all grid cells in sub-Saharan Africa over time. We do this by using all the DHS surveys to estimate the RF-2 model and then use

this model to predict the log of average consumption for each 10x10km cell in 42 countries from 2003-18. Poverty rates for each cell are calculated by applying the approach described in section 4.4. We refer to the resulting estimates as the "SED" data, for *Spatial Economic Development* data. As mentioned in the Introduction, dynamic versions of these maps covering the entire time period are found at https://www.spatial-economic-development.com/.

Figure 6 shows the population-weighted distribution of the cell estimates. For comparison purposes, we have also calculated the cell values using RF-1, which uses only nightlights as predictors. The histograms include all grid cells (over 4.1 million) across countries and over time. The left panel shows the distributions for log consumption. The top left histogram presents the consumption distribution resulting from RF-1. The distribution is concentrated on an extremely small range of values, with a huge proportion of values in the same narrow band. Using the SED data (bottom left panel), there is much more variability in the estimates and much higher maximum values. The poverty rates are depicted in the right panels, and again there is a concentration of (very high) values from RF-1, and a much more fine-grained distribution of poverty using the SED data. The histograms therefore clearly illustrate that nightlights alone cannot produce estimates with the variability we would expect in spatial measures of well-being, and that the inclusion of the additional predictors results in SED measures that exhibit the rich variability we should expect.

Does this rich variability accurately capture variation in economic-well being across cells within countries? While we cannot address this question at the cell level, we can examine whether within-country variation in SED estimates is aligned with variation in estimates from external datasets that exist at the level of subnational regions. The external data sets, along with the SED data, provide estimates with unknown biases and unknown measurement error. Thus, these comparisons should not be viewed as a validation exercise of the SED cell-level data. But since each dataset is computed using different data sources and different methodologies, if the SED and external data produce estimates that are strongly correlated within countries, this should improve our confidence in each data set, and in the possibility that the cell estimates from SED data are capturing within-country variation across cells in a meaningful way.

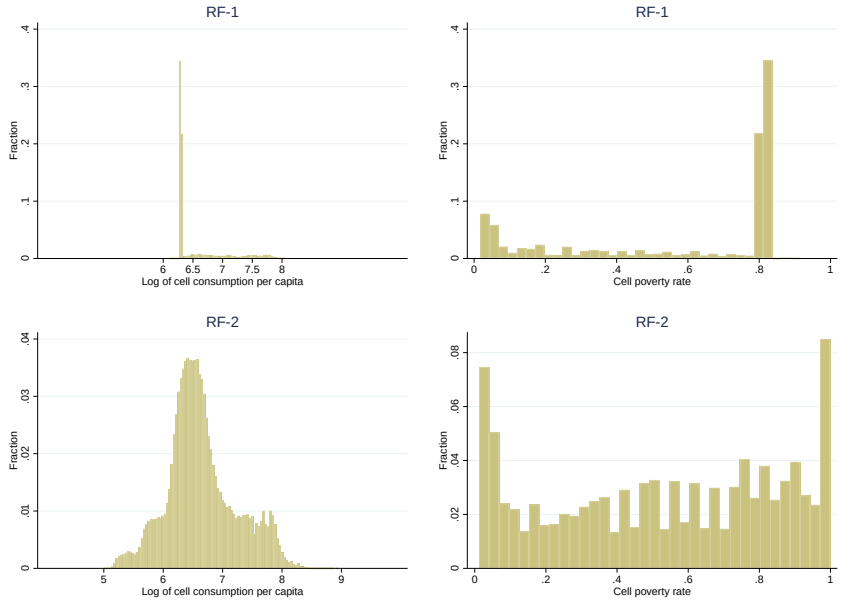


Figure 6. The population-weighted distribution of consumption and poverty cells in SED. The left panel histograms depict the population weighted distribution of consumption from RF-1 and RF-2. The right panel depicts these distributions for poverty rates. The data include over 4.1 million cells from 2003-18.

We begin by using data from the Human Development Index (HDI), which has been compiled for subnational regions by Smits and Permanyer (2019). The subnational HDI is constructed in the same way as the national HDI, and is an aggregate of three indices: health (measured as life expectancy at birth), education (based on the mean of average years of schooling and expected years of schooling), and standard of living (measured as log of gross net income (GNI) per capita in 2011 PPP dollars).[30] We are particularly interested in comparing the SED consumption measure with the the HDI measure of regional income per capita. Although income obviously measures something different than consumption, the two variables should be closely related, and the goal here is to understand how regional income is correlated with SED regional consumption within countries.

To make the comparison, we aggregate the SED measures of consumption in each cell to the level of the HDI regions.[31] The data are from 2016, a year which minimizes interpolation of the HDI data. There are a total of 244 matched regions from 20 countries.

[30] See Smits and Permanyer (2019) for a detailed discussion of the data sources.

[31] For HDI, we have only the name of the region. We therefore assign each grid cell to the region associated with administration level 1 (akin to a state, and closest to the regional level used in the external data). We then match to the extent possible the names of these regions to the names of the regions in HDI.

For regional poverty, we use the World Bank's GSAP estimates of poverty rates in subnational regions.[32] We can use SED poverty rates along with the population of each grid cell to calculate the poverty rate in corresponding subnational regions. This makes it possible to conduct within-country comparisons of poverty rates from the SED data with poverty rates from GSAP. GSAP estimates are based on survey data from the Global Monitoring Database (GMD) and exist only for 2019. Though the data are based on surveys, we do not have information on the representativeness of the surveys at the subnational level or on the number of respondents on which the regional estimates are based. The data are created by using one survey for each country, the one conducted closest to 2019. In some cases, the survey on which the data are based are from many years before, and the data are then interpolated to 2019 using national-level estimates of poverty change. To minimize errors associated with this interpolation, we focus only on countries where the baseline survey is within 5 years of the SED data we use, which is from 2018 (so we are comparing 2018 SED data with 2019 GSAP estimates). Our regional poverty data set has SED and GSAP estimates for 391 regions from 29 countries.

	HDI Income (1)	HDI Education (2)	HDI Life Exp. (3)	GSAP Poverty (4)
Consumption, SED	0.81	0.64	0.38	
Poverty, SED				0.67

Table 2. WITHIN-COUNTRY CORRELATION USING DATA AT THE LEVEL OF SUBNATIONAL REGIONS. The cells display within-country correlations between variables from external data (named at the top of each column) and variables from the SED data (named at the beginning of each row). The within-correlation is based on deviations of the variables from their country means. In columns (1) through (3), there are 244 regions from 20 countries. In column (4) there are 391 regions from 29 countries.

Table 2 provides the "within-country" correlations of SED with the external variables. These within-correlations are calculated using the deviations of each region from its country mean. Columns (1) through (3) of Table 2 present the correlations for the three components of the HDI using the 244 regions. We find that the SED estimate of regional consumption has a correlation of 0.80 with the HDI measure of income per capita. This is stronger than the

[32] For details regarding the construction of the GSAP data, see https://datacatalog.worldbank.org/search/dataset/0042041/global_subnational_poverty_atlas_gsap. Our data our from the 4th edition.

within-correlation for the other two components of the HDI. Column (4) of Table 2 presents the within-correlation for the poverty rates, which is again quite large, though smaller than that in column (1). Thus, the evidence in Table 2 indicates that the regional variables created by aggregating SED cells to the regional level are strongly correlated within countries to the two external data sources.

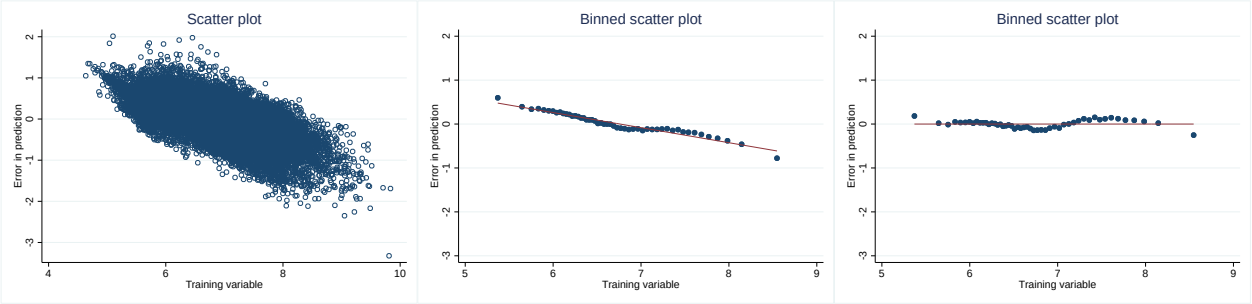
6. Using SED data in regression analysis: the problem of NCME

The SED dataset can be harnessed to pursue a variety of distinct objectives. One is to provide detailed, ground-level descriptions of consumption and poverty levels, which can assist in addressing humanitarian needs by guiding the geographic targeting of development and humanitarian relief projects. A second important and distinct application is to use the SED estimates as proxies for economic well-being in regressions using a spatially disaggregated unit of analysis. This will be our focus in what follows.

For the reasons described at length in section 2, any data used in place of NL in regression analysis should not only be accurate, it should should also be free of the nonclassical measurement error that plagues NL. As noted above, however, we should not expect SED to be free of such error. Panel (a) in Figure 7 shows the relationship between the prediction error (from RF-2) and the training variable. This graph highlights a negative correlation between these variables, indicating that SED tends to over-predict cells with lower income and under-predict cells with higher income. As a result, using this data in regression analysis can lead to biased coefficients for the reasons discussed in section 2. What follows is a simple adjustment to address this problem.[33] This proposed adjustment can always be applied to proxies generated via supervised machine learning when it is reasonable to assume the training variable is a representative sample of the target variable.

Our approach uses the training variable as an auxiliary sample, allowing us to compute a simple transformation of SED that is free from nonclassical error. It works as follows. Consider a target variable y for which a proxy $\tilde{y}$ exists that might contain non-classical

[33]Several estimators have been proposed to address nonclassical measurement error (see Chen et al. (2011) for a review). While these techniques could be applied when using the SED data, they would be unnecessary if we could rid the SED data of the NCME, thus making it possible to use conventional estimation methods when using the adjusted SED data.



(a) Prediction error (RF-2) vs. training data: scatterplot

(b) Prediction error (RF-2) vs. training data: binned scatterplot

(c) Prediction error (RF-2) from transformed data, vs. training data: binned scatterplot

Figure 7. Prediction error versus Training variable. Panel (a) plots the prediction error from $\hat{y}_{RF2}$ versus the training variable; Panel (b) presents the binned scatter plot of the same relationship; Panel (c) displays the binned scatter plot relating the prediction error from $\hat{y}^T_{RF2}$ versus the training variable.

measurement error, ν. In our context, y and $\tilde{y}$ represent consumption per capita and its prediction, respectively. We consider the linear projection of $\tilde{y}$ on y,[34]

$$\tilde{y} = \alpha_0 + \alpha_1 y + \epsilon. \tag{12}$$

By definition of linear projection, ϵ and y are uncorrelated. This allows us to define a new proxy $\tilde{y}^T$ as

$$\tilde{y}^T = \frac{\tilde{y} - \alpha_0}{\alpha_1} = y + \epsilon/\alpha_1. \tag{13}$$

The new proxy variable $\tilde{y}^T$ and $\tilde{y}$ each have the same correlation with y, but since y and ϵ are uncorrelated, $\tilde{y}^T$ contains only classical measurement error.

To estimate α_0 and α_1, we assume that the training variable provides a representative sample of y. The training data can therefore be used to derive consistent estimates of these parameters. We do this by regressing the predictions (from model RF-2) obtained in section 4 on the training variable to obtain $\hat{\alpha}_0$ and $\hat{\alpha}_1$, thus allowing us to compute $\hat{y}^T$. Panel (c) of Figure 7 displays the binned scatter plot of the prediction error associated with $\hat{y}^T$ versus the training variable. It confirms there is no relationship between these variables.

Table 3 provides summary statistics of the training variable, the prediction of consumption from RF-2, denoted as $\hat{y}_{RF2}$, and the transformed prediction, $\hat{y}^T_{RF2}$. Columns 1-4 of Table 3

[34] Panel (b) in Figure 7 suggests that the relationship between u and the training variable is in fact linear in our case, implying that this linear equation captures well the relationship between these variables.

display the mean, standard deviation and min and max values for the three variables. $\hat{y}_{RF2}$ has less variability than the training variable while $\hat{y}^T_{RF2}$ has more variability, but otherwise the values are quite similar. Column 5 reports correlations between the prediction errors and the training variable: as expected, the correlation between the error in $\hat{y}^T_{RF2}$ and the training variable is zero. Column 6 shows that both proxies for consumption have a large and identical correlation with the training variable (as one is just a linear transformation of the other). Finally, column 7 shows that $\hat{y}_{RF2}$ has less prediction error than $\hat{y}^T_{RF2}$ (the MSE of $\hat{y}^T_{RF2}$ is around 50% larger). The table therefore illustrates a tradeoff between $\hat{y}_{RF2}$ and $\hat{y}^T_{RF2}$: while the former is more accurate and therefore potentially better for descriptive purposes, the latter does not have the nonclassical measurement error that contaminates $\hat{y}_{RF2}$ and therefore is more suitable in regression analysis.

	Mean	Std	min	max	$corr_{(\hat{\varepsilon},\widehat{y^*_{rct}})}$	$corr_{(\hat{y},\widehat{y^*_{rct}})}$	MSE
	[1]	[2]	[3]	4	5	6	7
$\widehat{y^*}$	6.74	0.699	4.634	9.835			
$\hat{y}_{RF2}$	6.74	0.558	5.284	8.888	-0.603	0.824	0.135
$\hat{y}^T_{RF2}$	6.74	0.849	4.512	9.998	0.000	0.824	0.183

Table 3. Transformed and Untransformed proxies of Consumption per capita. This table presents summary statistics of the training variable ($\widehat{y^*}$), its best predictor ($\hat{y}_{RF2}$) and the transformed best predictor ($\hat{y}^T_{RF2}$). The latter is computed as $\hat{y}^T_{RF2} = (\hat{y}_{RF2} - 2.32)/.657$, where these coefficients have been computed in a regression of $\hat{y}_{RF2}$ on $\widehat{y^*_{rct}}$.

7. Illustration: Institutions and economic development

This section illustrates the usefulness of the SED data by revisiting two influential papers in the literature on institutions and economic development, Michalopoulos and Papaioannou (2013) and (2014) ("MP13" and "MP14" in what follows). Both papers use NL as a dependent variable to study the effects of institutions on development. To gain insight into how the NCME in NL affects their conclusions, we will estimate models from MP13 and MP14 using $\hat{y}^T_{RF2}$ in place of NL.

The usefulness of the SED data in this context manifests in two significant ways. First, SED estimates of consumption allow us to obtain more accurate estimators, free from the biases associated with the use of NL. Second, since the SED data is measured in consumption dollars,

its opens up the possibility of not just identifying the presence of significant relationships but also of quantifying the extent to which institutional variables contribute to development. This capability is crucial for determining what truly drives development, allowing researchers and policymakers to differentiate between statistically significant results and those that have substantial, practical implications in the real world.

In the remainder of this section, we begin by substituting $\hat{y}^T_{RF2}$ for NL in the analyses of MP13 and MP14. The resulting conclusions about the role of institutions in development are markedly different than those found in the original MP papers. We then draw on the discussion in section 2 (regarding biases stemming from non-classical measurement error in NL) to explain why the resulting conclusions change so much when we substitute $\hat{y}^T_{RF2}$ for NL.

7.1. Re-estimating models of institutions and development using SED. MP13 and MP14 use a dataset with a common general structure, one that relies on the geographically fine-grained feature of nightlights data. The unit of analysis is a pixel of 0.125 X 0.125 decimal degrees (approximately 12.5 km X 12.5 km).[35] Each pixel is assigned to an ethnic group based on maps of ethnic homelands described in Murdock (1967), and values of nightlights in the pixels are used as a proxy for economic development. To replicate the MP13 and MP14 analyses with the SED data, we use the value of $\hat{y}^T_{RF2}$ that is closest to each MP pixel. The average distance between the pairs of centroids is 3.7 km in both MP13 and MP14. The MP data include all of Africa, whereas the SED data includes sub-Saharan African countries, which means we do not have data for five countries in the MP data.[36] MP13 and MP14 present a wide range of models in an attempt to assess the robustness of results regarding institutions, and here we present illustrative results for one table from each paper. Section F in the Appendix presents results from additional empirical models.

7.1.1. The role of national institutions. First consider MP14, which finds no direct relationship between national institutions and nightlights. A core set of pixel-level results are found in Panel B of their Table IV (p. 177) and, for convenience, they are reproduced in Panel A of Table 4. These models exploit the fact that ethnic group boundaries often transcend national

[35]The MP papers also include ethnic group-level analyses, which are not our focus here.
[36]These are Algeria, Egypt, Libya, Morocco and Tunisia.

borders and that in the case of Africa, these borders can be considered as exogenous. Thus, by considering only partioned groups (whose settlement areas include both sides of a national border) when estimating a model that includes for each pixel a measure of the quality of national institutions, it is possible to compare the value of nightlights for the same ethnic group across the boundaries of countries with different values of the institutional variables. The dependent variable in Panel A of Table 4 is an indicator that takes the value 1 if the pixel is "lit" (i.e., not completely dark), and the World Bank's measures of RULE OF LAW and CONTROL OF CORRUPTION are the national-level measures of institutions. The first four columns in Panel A use RULE OF LAW as the measure of national institutions and the last four columns use CONTROL OF CORRUPTION. The key finding is that in models without ethnic-group fixed effects (columns 1, 3, 5 and 7) the coefficients for RULE OF LAW and CORRUPTION are positive and significant, but once ethnic-group fixed effects are added (columns 2, 4, 6 and 8), there is no precisely estimated effect of RULE OF LAW or CONTROL OF CORRUPTION on nightlights.

Panel B of Table 4 re-estimates the same models as Panel A using the subset of pixels that are common to the SED and MP14 data (i.e., pixels from the five north African countries are excluded, which amounts to around 4% of the total number of cells). Results are very similar, but now the estimated coefficients of the institutions variables in regressions including ethnic fixed effects are a bit larger and marginally significant. As discussed below, the marginally significant results for the institutions variables in Panel B are not robust to other specifications considered in MP14.

Panel C re-estimates the same models using the measure of log consumption from the full SED model, $\hat{y}^{T}_{RF2}$. This model produces results that are very supportive of arguments about national institutions. All coefficients in the fixed effects regressions are now significant at the 5% level and the overall fit of the model improves tremendously.[37] Importantly, we can now provide a meaningful estimate of the magnitude of the effect of national institutions on development, something that is challenging to do when NL is the outcome. RULE OF LAW is a

[37]The difference between the results using nightlights and results using the SED data *is not* explained by the way the training variable is constructed. To see this, we have also estimated the models in the table using log consumption as estimated by RF-1 (the model that uses only NL as predictors, but which obviously has the same training variable as RF-2). The results are very similar to those presented in Panels A and B, with less precisely estimated coefficients and coefficients of quite modest size.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Panel A: Dep. variable is Nightlights (MP14, Original Sample)							
RULE OF LAW	0.1072***	0.0246	0.0834**	0.0278				
	(0.0400)	(0.0165)	(0.0324)	(0.0181)				
CONTROL OF CORRUPTION					0.1371***	0.0370	0.1097***	0.0403
					(0.0464)	(0.0273)	(0.0415)	(0.0290)
Adj. R-squared	0.149	0.331	0.202	0.340	0.160	0.331	0.209	0.340
N	42710	42710	41025	41025	42710	42710	41025	41025
	Panel B: Dep. variable is Nightlights (Reduced Sample)							
RULE OF LAW	0.0850**	0.0311*	0.0759**	0.0370*				
	(0.0428)	(0.0170)	(0.0369)	(0.0199)				
CONTROL OF CORRUPTION					0.1121**	0.0479*	0.1025**	0.0541*
					(0.0523)	(0.0271)	(0.0482)	(0.0296)
Adj. R-squared	0.131	0.262	0.149	0.271	0.140	0.262	0.156	0.271
N	40872	40872	39251	39251	40872	40872	39251	39251
	Panel C: Dep. variable is log consumption p.c., model RF-2 (Reduced Sample)							
RULE OF LAW	0.4927**	0.2377**	0.3617***	0.1641**				
	(0.1979)	(0.0942)	(0.1222)	(0.0717)				
CONTROL OF CORRUPTION					0.6509***	0.3162***	0.4884***	0.2593***
					(0.1792)	(0.1144)	(0.1136)	(0.0872)
Adj. R-squared	0.219	0.784	0.434	0.812	0.297	0.785	0.474	0.816
N	40872	40872	39251	39251	40872	40872	39251	39251
Ethnicity fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
Population density and area	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location controls.	No	No	Yes	Yes	No	No	Yes	Yes
Geographic controls	No	No	Yes	Yes	No	No	Yes	Yes

Table 4. NATIONAL INSTITUTIONS AND ECONOMIC DEVELOPMENT. This table re-estimates models in Table IV, panel B (which are at the pixel level) of Michalopoulos and Papaioannou (2014). The coefficients are from OLS models with double-clustered standard errors in parentheses. See Michalopoulos and Papaioannou (2014) for full details regarding the data and estimation. The results presented in Panels B and C are for a subset of MP data, as the RF data do not include five north African countries. * indicates p<.10, **indicates p<.05, and *** indicates p<.01.

continuous variable with a range of 2.8475. Therefore, using the estimates from column (4), going from the worst to the best value of RULE OF LAW implies a 46% increase in consumption per capita. Similarly, going from the worst to the best value of CONTROL OF CORRUPTION implies a 64% increase in consumption per capita (using the estimates from column (8)).[38] To put a value on these very large effects, the Democratic Republic of Congo has a RULE OF LAW score of -1.88 and Botswana has a RULE OF LAW score of 0.615. Thus, if the DRC could improve its rule of law institutions to the same quality as Botswana, this would result in a 41% increase in consumption per capita (using model 4). In 2019 (the most recent year for which data is

[38] A one standard deviation increase in RULE OF LAW implies an 9.6% increase in consumption and a one standard-deviation in CONTROL OF CORRUPTION implies a 15.2% increase in consumption.

available), the consumption per capita in DRC was $633, so an improvement in institutions of this magnitude would net each person in DRC $260.

Section F in the Appendix provides results from estimating other models in MP14 using nightlights and the SED estimates. Across all models that have relevant fixed effects, the coefficient on institutions is never significant (even at the 10% level) when nightlights is the dependent variable. By contrast, when consumption is the dependent variable, the coefficients on the national institutions variables are in general significant at the 5% level.[39]

In sum, the evidence here indicates that the previously observed weak relationship between national institutions and development is a consequence of using nightlights as the development proxy. The conclusions shift dramatically when SED consumption is used as the measure of development, revealing that institutions, especially the rule of law, have a large positive impact on development.

7.1.2. The role of ethnic institutions. Next consider MP13, which finds a robust relationship between the centralization of ethnic institutions and nightlights. The models in this paper exploit boundaries between ethnic groups within countries. The goal is to estimate whether development (measured using NL) is higher in pixels that are in areas with more centralized pre-colonial ethnic institutions. A core set of pixel-level results are found in Panel A of Table V of MP13 that, for convenience, we produce in Table 5, Panel A. The dependent variable in the first five columns of their table, our focus here, is a dichotomous measure of nightlights,[40] and the measure of pre-colonial institutions, JURISDICTIONAL HIERARCHY, is Murdock's (1967) index of "Jurisdictional Hierarchy beyond the local community level." This discrete variable ranges from 0 to 4.[41] The main result is that JURISDICTIONAL HIERARCHY is positive and significant at the 5% level.

[39] Specifically, we use both NL and $\hat{y}^T_{RF2}$ to re-estimate the pixel-level models from MP14 Table V (which focuses on pixels that are close to a national border), Table VI (which adopts a spatial regression discontinuity approach by adding RD-polynomials based on the distance of a pixel to the border) and Table VII (which has RD models similar to those in Table VI but focuses only on large institutional differences across borders). The only exception to significant coefficients using SED data is the results for CONTROL OF CORRUPTION in the models from MP14 Table VI.

[40] We also estimate the the models with the log of lights, as MP13 do in models 6-10, and the results are the same (see section F in the Appendix).

[41] A zero score indicates stateless societies, a value of 1 corresponds to petty chiefdoms, 2 designates paramount chiefdoms, while 3 and 4 indicate groups that were part of large states.

	(1)	(2)	(3)	(4)	(5)
	Panel A: Dep. var. is lit/unlit (MP13, Original Sample)				
JURISDICTIONAL HIERARCHY	0.0673**	0.0447**	0.0280***	0.0308***	0.0265***
	(0.0314)	(0.0176)	(0.0081)	(0.0074)	(0.0071)
Adj. R-squared	0.034	0.272	0.358	0.3757	0.379
N	66570	66570	66570	66173	66173
	Panel B: Dep. var. is lit/unlit (Reduced Sample)				
JURISDICTIONAL HIERARCHY	0.0301	0.0349*	0.0238***	0.0256***	0.0173***
	(0.0203)	(0.0178)	(0.0088)	(0.0088)	(0.0060)
Adj. R-squared	0.008	0.182	0.268	0.287	0.293
N	61359	61359	61359	61015	61015
	Panel C: Dep. var. is log consump. p.c., RF-2 (Reduced Sample)				
JURISDICTIONAL HIERARCHY	-0.0089	-0.0082	-0.0184	-0.0150	-0.0151
	(0.0625)	(0.0234)	(0.0233)	(0.0177)	(0.0181)
Adj. R-squared	0.000	0.777	0.792	0.831	0.836
N	61359	61359	61359	61015	61015
Country Fixed effects	No	Yes	Yes	Yes	Yes
Population Density	No	No	Yes	Yes	Yes
Controls at the Pixel level	No	No	No	Yes	Yes
Controls at the Ethnic-Country level	No	No	No	No	Yes

Table 5. ETHNIC INSTITUTIONS AND ECONOMIC DEVELOPMENT. This table re-estimates models in Table V, panel A, models 1-5 (which are at the pixel level) of Michalopoulos and Papaioannou (2013). The coefficients are from OLS models with double-clustered standard errors in parentheses. See Michalopoulos and Papaioannou (2013) for full details regarding the data and estimation. The results presented in Panels B and C are for a subset of MP data, as the SED data do not include five north African countries that are included in MP13. Each panel in the table differs only in the dependent variable used to measure economic well-being. ***, **, and * indicate statistical significance at the 1%, 5% and 10% levels, respectively.

Panel B of Table 5 re-estimates the models using the subset of pixels that are common to the SED and MP13 data (i.e., the five North African countries are excluded). The results are in broad agreement with those in Panel A. Panel C uses SED log-consumption p.c. (RF-2) as the outcome variable, and the results are very different: the coefficients for JURISDICTIONAL HIERARCHY are now insignificant in all columns and have the opposite sign. In section F of the Appendix, we present the results from a wide range of MP13 models, and the central finding is that there is no relationship between JURISDICTIONAL HIERARCHY and the SED measures of consumption.

7.2. Non-classical measurement error and the divergent results. The divergent results across panels in Table 4 and in Table 5 are striking. The original results based on nightlights suggest that centralized ethnic institutions matter for development whereas strong national institutions do not. The results here suggest the opposite. If the results obtained using SED

accurately reveal the true relationship between institutions and development, the findings in MP14 would be subject to *attenuation* bias, which could cause a nonexistent relationship to be identified when in fact there is a positive relationship. This implies a *negative* bias in MP14 estimates. Conversely, the results in MP13 would exhibit *amplification* bias, where a positive relationship is reported even though in reality, none exists. This implies a *positive* bias in MP13. Are the directions of these biases consistent with the presence of NCME in NL?

As described in section 2, when nightlights are used as a dependent variable, a bias δ arises whenever the regressors are correlated with the error in NL, u, and the direction of this bias depends on the sign of that correlation. In the MP14 case, since the posited bias is negative, in order for this bias to be due to NCME, national institutions must be negatively correlated with the error u. In the MP13 case, since the posited bias is positive, in order for this bias to be due to NCME, centralized ethnic institutions must be positively correlated with the error in nightlights. We examine whether these correlations are likely to be present in the data.

Consider first the national institutions (NI) results in MP14. If the NI variables are in fact correlated with development, as Panel C in Table 4 strongly suggests, what would be the sign of the bias δ in MP14 regressions? Recall that u and y^* (the "true" measure of development) are negatively related as discussed in section 2. Since the Panel C results indicate that y^* and the NI variables are positively related, then a negative relationship between NI and u must follow. This implies that if the true relationship between NI and development is in fact positive, estimates obtained using NL will be attenuated, i.e., biased towards zero, which is consistent with our findings using an outcome variable that does not have NCME.

Consider next the MP13 results on the relationship between JURISDICTIONAL HIERARCHY (JH) and development. The discussion in section 2 reminds us that if JH and u are positively correlated, then a positive relationship between NL and JH can be found *even in situations where no relationship in fact exists*. There is strong reason to believe that this positive relationship between JH and u should exist because of the relationships between JH, NL and population density.

First consider the relationship between the error in NL, u, and population density. It is well-understood that nightlights detected by satellites are overwhelmingly in places with high population density, such as urban areas (e.g., Gibson et al, 2020). This implies that "false negatives" ($u = -1$) will be more likely in areas with low concentration of people, and "false positives" ($u = 1$) will be more likely in the opposite case. Thus, there should be a positive correlation between population density and u.

Next consider the relationship between JH and population density. It is well-established that population density is related to the development of institutions. Historically, as populations grew, social organisation became more complex, creating a greater need for coordination in decision-making (see, for instance, Turchin et al, 2022). JURISDICTIONAL HIERARCHY aims to capture precisely the nature of institutions that made such coordination possible. In the MP13 data, the correlation between population density and JURISDICTIONAL HIERARCHY is indeed positive and highly significant.[42]

Since JH is positively related to population density and population density is positively related to u, we should expect a positive relationship between JURISDICTIONAL HIERARCHY and u, which will lead to an upward bias that amplifies the estimated effects of JURISDICTIONAL HIERARCHY in nightlights models.[43] As a result, this amplification bias can lead to findings of positive effects even in situations where there is no relationship between JH and development.

In sum, our analysis of two research papers utilizing nightlights highlights a crucial point: biases are likely to occur and they can be either positive or negative. Indeed, in the two papers examined here, the biases from NL models seem to lead to conclusions about institutions that are the opposite of those reached when using a measure that has no NCME. The challenge of drawing definitive conclusions from nightlights-based models therefore is deep and intractable.

[42]More specifically, the correlation between (the log of) population density and JH is 0.15, which is more than 50% larger than the 0.09 correlation between nightlights and JURISDICTIONAL HIERARCHY, which is the central relationship of interest in MP13.

[43]It is important to note that linearly controlling for population density may attenuate but not eliminate this problem of bias. If JH and u are non-linear functions of population density, then linearly controlling for the latter will not eliminate the bias completely. But the relationship between JURISDICTIONAL HIERARCHY (or u) and population density *cannot* be exactly linear, as u and JURISDICTIONAL HIERARCHY are discrete variables whereas population density is continuous. This implies that introducing population density in the nightlights regressions can reduce the positive bias of the coefficient for JURISDICTIONAL HIERARCHY (as we see in column 3 across panels in Table 5), but it is unlikely to solve the bias problem described in this paragraph.

8. Conclusion

Though widely used, nightlights are a problematic proxy for economic development in regression analysis because of non-classical measurement error. We have shown that this measurement error can lead to biased coefficients in research that uses nightlights as a spatially disaggregated proxy for development, and that those estimated coefficients can attenuate or amplify the true relationship between the relevant variables. The measures we have developed in its place not only avoid the biases inherent in the use of nightlights, they make it possible to interpret empirical results in a substantively meaningful way. Our results suggest, for example, that contrary to findings using nightlights, the cells in a country with the highest level of control of corruption can be expected to have average consumption levels that are 64% higher than that of cells in countries with the lowest level of control of corruption. Thus, the substantively interpretable metric of the SED data, along with the high level of spatial granularity, can be used to gain a more nuanced understanding of economic development. This is of value not only to researchers, but also to policy markers seeking to identify specific areas of progress or specific areas requiring targeted interventions. Because of its simplicity and low computational cost, the framework described here is easily scalable to contexts beyond Africa, facilitating comparative studies and broadening the scope of empirical research on economic development.

While we believe the limitations of nightlights for spatially disaggregated research and policymaking are clear, an important avenue for future research involves understanding the limitations of nightlights when they are aggregated at high levels, such as at the national level. For example, Martinez (2022) aggregates nightlights at the national level to show that dictators exaggerate economic development. Whether the non-classical measurement error in nightlights emphasized here is problematic at such a high level of aggregation is not obvious, and exploring this issue will deepen our understanding of the limitations of nightlights.

9. References

Aiken, E., Bellue, S., Karlan, D., C. Udry and J.E. Blumenstock (2022), "Machine learning and phone data can improve targeting of humanitarian aid". *Nature* 603, 864–870.

Bound J., C. Brown, G. J. Duncan, and W.L. Rodgers (1994), "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Dat," *Journal of Labor Economics*, 12, 345-368.

Burke, M., A. Driscoll, D. Lobell, and S. Ermon (2021), "Using satellite imagery to understand and promote sustainable development," *Science* 371.

Breiman, L. (2001), "Random Forests", *Machine Learning*, 45, 5–32.

Chen, X., and W.D Nordhaus (2011), "Using luminosity data as a proxy for economic statistics," *Proceedings of the National Academy of Sciences*, 108, 8589–8594.

Chi, G, S. Fang, and J. E. Blumenstock, (2022). "Microestimates of Wealth and Poverty for all Low- and Middle-Income Countries," *Proceedings of the National Academy of Sciences*, 119(3), 1-11.

Duclos, J., J.Esteban and D. Ray (2004), "Polarization: Concepts, Measurement, Estimation," *Econometrica* 72, 1737-1772.

Gibson, J., O. Susan, and G. Boe-Gibson (2020), "Night Lights in Economics: Sources and Uses," *Journal of Economic Surveys* 34, 955–980.

Henderson, J. V., A. Storeygard, and D. Weil (2012), "Measuring Economic Growth from Outer Space," *American Economic Review* 102, 994–1028.

Henderson, J. V., A. Storeygard, and D. Weil (2011), "A Bright Idea for Measuring Economic Growth," *American Economic Review* 101, 194–199.

Hruschka, D.J. , D. Gerkey and C. Hadley (2015) "Estimating the absolute wealth of households," *Bulletin of the World Health Organization* 93, 483–490.

Jean, N., M. Burke, M. Xie, M. Davis, W. Matthew, D. B. Lobell, and S. Ermon (2016), "Combining satellite imagery and machine learning to predict poverty," *Science* 353, 790–794.

McCallum, I., Kyba, C.C.M., Bayas, J.C.L., E. Moltchanova, M. Cooper, J. Crespo Cuaresma, S. Pachauri, L. See, O. Danylo, I. Moorthy, M. Lesiv, K. Baugh, C. D. Elvidge, Martin Hofer and S. Fritz (2022), "Estimating global economic well-being with unlit settlements." it Nature Communications 13, 2459.

Martinez, L.R (2022), "How Much Should We Trust the Dictator's GDP Growth Estimates?," *Journal of Political Economy*. 130

Meyer, B. D, and N. Mittag (2017) "Misclassification in binary choice models," *Journal of Econometrics* 200, 295-311.

Michalopoulos, S., and E. Papaioannou (2013), "Pre-Colonial Ethnic Institutions and Contemporary African Development," *Econometrica* 81, 113–152.

Michalopoulos, S., and E. Papaioannou (2014), "National Institutions and Subnational Development in Africa," *The Quarterly Journal of Economics* 129, 151–213.

Michalopoulos, S., and E. Papaioannou (2018) "Spatial Patterns of Development," *Annual Reviews* 10, 383-410.

Murdock, G.P. (1967), "Ethnographic Atlas: A Summary," *Ethnology* 6, 109.

Pinkovskiy, M., and X. Sala-i-Martin (2016), "Lights, Camera ... Income! Illuminating the National Accounts-Household Surveys Debate," *The Quarterly Journal of Economics* 131, 579–631.

Scholao, M. and R.Y. Zou (2020), "The random forest algorithm for statistical learning," *The Stata Journal* 20, Issue 1.

Smits, J., and I. Permanyer (2019), "The Subnational Human Development Database," *Scientific Data* 6, 190038.

Turchin, P., H. Whitehouse, S. Gavrilets, D. Hoyer, P. François, J. Bennett, K. Feeney, P. Peregrine, G. Feinman, A. Korotayev, N. Kradin, J. Levine, J. Reddish, E. Cioni, Enrico, R. Wacziarg, G. Mendel-Gleason, M. Benam, (2022) "Disentangling the evolutionary drivers of social complexity: A comprehensive test of hypotheses," *Science Advances*, 8 eabn3517.

Yeh, C., A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke (2020) "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa," *Nature Communications*, 11, 2583.

Yeh, C., S. Meng, A. Wang, E. Driscoll, P. Rozi, J. Liu, J. Lee, M.Burke, D. B. Lobell, S. Ermon (2021), "SustainBench: Benchmarks for Monitoring the Sustainable Development Goals with Machine Learning',' Thirty-fifth Conference on Neural Information Processing Systems, Datasets and Benchmarks Track (Round 2).

Appendix (For Online Publication)

This appendix is divided into six sections:

- Section A shows a figure documenting the increase in the use of nightlights in economics over time.
- Section B lists the DHS surveys used in constructing the training variable, and presents further details about its construction.
- Section C presents a table that describes the variables used in the prediction exercise.
- Section D provides additional information about the prediction models, including parameter tuning, variable importance in the prediction models, prediction accuracy, and a comparison of the prediction results from our RF models with those of Yeh et al. (2020) (section D.5).
- Section E provides further information about the approach we follow to estimate cell-level poverty rates.
- Section F provides additional information and estimation results regarding the analysis of the models in Michalopoulos and Papaioannou (2013 and 2014) using the SED data.

A. Nightlights in Economic Research

Figure A.1 displays the evolution of the number of papers referencing "nightlights" in economics according to Google Scholar. For additional information about the use of nightlights over time in economics research, see Gibson, et al (2020).

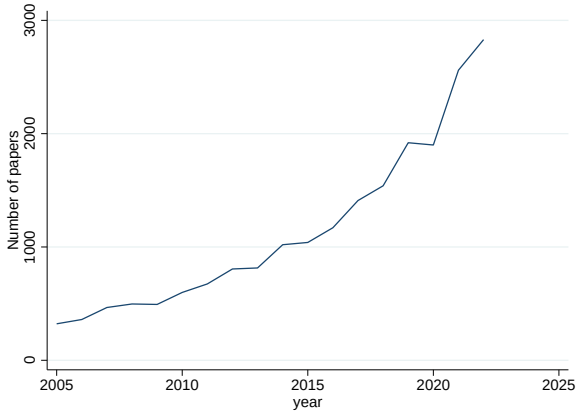


Figure A.1. Papers in economics referencing nightlights. Source: Google Scholar. The graph depicts the number of papers in Google Scholar obtained using the keywords "nighlights+economics" from 2005 to 2022.

B. Constructing the training variable, $\widehat{y^*_{rct}}$.

This section provides additional details about the construction of the training variable, $\widehat{y^*_{rct}}$. Here we summarize the main steps involved in the process; the subsequent sections provide additional details related to each of these steps.

(1) For each survey, estimate a principal components model using all respondents and a variety of asset variables that are present in the given survey to generate the asset measure, y^A_{ict}.

(2) Use the WB-PIP data on average consumption and the Gini coefficient for the country-year of the survey to obtain estimates of μ^C_{ct} and σ^C_{ct}, the mean and the standard deviation of log-consumption measured in 2011 PPP dollars.

(3) Apply the estimates of y^A_{ict}, μ^C_{ct} and σ^C_{ct} to compute $\widehat{y^*_{ict}}$ as described in section 3. The resulting $\widehat{y^*_{ict}}$ is expressed in log of consumption in 2011 PPP dollars.

(4) Calculate $\widehat{y^*_{rct}}$ by averaging $\widehat{y^*_{ict}}$ for all i in cluster r.

B.1. Constructing the asset index, y^A_{ict}. We use 85 DHS surveys comprising over 900,000 households who are sampled from across 29 sub-Saharan African countries in the period 2006-2018. Figure B.1 displays a map of the African countries for which we have DHS data, and Table B.1 provides a list of all the DHS surveys used to create the training data. DHS surveys provide information on household-level ownership of different assets. These asset variables are related to sanitation in the home (the source of drinking water and type of toilet facility), the nature of the household's dwelling (flooring, wall, and roof materials; presence of electricity and number of sleeping rooms), and the presence of particular assets (e.g., radio, television, refrigerator, motorcycle or scooter, car or truck, telephone and mobile phone).

For each individual survey, we use principal component analysis (PCA) to create an asset index. The (unrotated) loadings of the asset variables on the first component are used to predict an aggregate asset score for each household. Since we estimate the PCA separately for each survey, the loadings vary across surveys. That is, we do not assume that the relationship of each asset to consumption is the same across time and space, which provides additional flexibility in constructing the index. Since DHS data is defined at the household level, we follow Duclos et al (2004) and divide the resulting index by $s^{0.5}$ to obtain a per-capita index, where s is the size of the household.[1] Finally, to calculate y^A_{ict}, we take the log of the index and standardize it so that for each survey the index has a mean of 0 and a standard deviation of 1.

B.2. Estimates of $\mu_{y^C_{ct}}$ and $\sigma_{y^C_{ct}}$. WB-PIP data, which are based on surveys rather than national accounts, provide the current gold standard for country level estimates of consumption in sub-Saharan Africa. WB-PIP provides estimates of the mean of consumption per capita but not of the mean and variance of its log, $\sigma_{y^C_{ct}}$ or $\mu_{y^C_{ct}}$. To overcome this limitation, we assume that y^C_{ict} follows a normal distribution with mean $\mu_{y^*_{ct}}$ and variance $\sigma^2_{y^C_{ct}} = \sigma^2_{y^*_{ct}} + \sigma^2_{\epsilon_{ct}}$, (see equation (3)). This implies that consumption per capita (the exponential value of y_{ict}, which we denote by x_{ict}) follows a log-normal distribution.[2] The log-normality assumption makes it possible to use estimates of the Gini coefficient, which exist in WB-PIP country-level data (as opposed to estimates of the variance of the distribution, which do not), to estimate the

[1] All the analysis in the paper has also been done without applying this correction and the results are virtually identical.

[2] See Battistin, Blundell and Lewbel (2009) for recent evidence supporting this claim.

Figure B.1. Countries in the training sample

standard deviation of log consumption. Under log-normality, the Gini coefficient associated with the consumption level, x_{ct}^C, is related to the standard deviation of the log of consumption, y_{ct}^C as follows:

$$\text{GINI}_{x_{ct}^C} = 2\Phi(\frac{\sigma_{y_{ct}^C}}{\sqrt{2}}) - 1, \tag{b.1}$$

where $\Phi(.)$ is the cumulative standard normal distribution. Therefore, the log-normality assumption allows us to estimate the variance of log-consumption, $\sigma^2_{y_{ct}^C}$, from estimates of the Gini of total consumption, which are typically available in WB-PIP. The log-normality assumption of the WB-PIP data also makes it straightforward to estimate $\mu_{y_{ct}^C}$ because the mean of total consumption, x_{ct}^C, and its log are related through the following equation:

$$\mu_{x_{ct}^C} = e^{\left(\mu_{y_{ct}^C} + 0.5(\sigma^2_{y_{ct}^C})\right)}.$$

Country	Year	Country	Year
Angola	2006	Mozambique	2009
Angola	2011	Mozambique	2011
Angola	2016	Mozambique	2015
Benin	2012	Mozambique	2018
Benin	2017	Namibia	2007
Burkina Faso	2010	Namibia	2013
Burkina Faso	2014	Nigeria	2008
Burkina Faso	2018	Nigeria	2010
Burundi	2010	Nigeria	2013
Burundi	2012	Nigeria	2015
Burundi	2016	Nigeria	2018
Cameroon	2011	Rwanda	2008
Chad	2015	Rwanda	2010
Dem. Rep. of Congo	2007	Rwanda	2015
Dem. Rep. of Congo	2013	Senegal	2008
Ethiopia	2011	Senegal	2011
Ethiopia	2016	Senegal	2013
Gabon	2016	Senegal	2014
Ghana	2008	Senegal	2015
Ghana	2014	Senegal	2016
Ghana	2016	Sierra Leone	2013
Guinea	2012	Sierra Leone	2016
Guinea	2018	Tanzania	2007
Kenya	2009	Tanzania	2010
Kenya	2014	Tanzania	2012
Kenya	2015	Tanzania	2015
Lesotho	2009	Tanzania	2017
Lesotho	2014	Togo	2014
Liberia	2007	Togo	2017
Liberia	2009	Uganda	2006
Liberia	2011	Uganda	2009
Liberia	2013	Uganda	2011
Liberia	2016	Uganda	2014
Madagascar	2011	Uganda	2016
Madagascar	2013	Uganda	2018
Madagascar	2016	Zambia	2007
Malawi	2010	Zambia	2013
Malawi	2012	Zambia	2018
Malawi	2014	Zimbabwe	2010
Malawi	2015	Zimbabwe	2015
Malawi	2017		
Mali	2006		
Mali	2012		
Mali	2015		
Mali	2018		

Table B.1. DHS SURVEYS. This table summarizes the DHS surveys employed in the construction of the training variable.

Finally, we apply the estimates of y^A_{ict}, μ^C_{ct} and σ^C_{ct} to compute $\widehat{y^*_{ict}}$ as described in section 3.1. The resulting $\widehat{y^*_{ict}}$ is expressed in log of consumption in 2011 PPP dollars.

A central challenge when using asset variables to compute a measure of economic well-being is achieving comparability across time and space. It is useful to contrast the way this comparability is achieved when using $\widehat{y^*_{ict}}$ with the approach in Yeh et al. (2020).[3] Yeh et al. (2020) achieve comparability in their asset-based variable by pooling all DHS surveys,

[3]Chi et al. (2022) also compute a training variable based on a DHS asset index, but this training variable only provides within country-year information, i.e., its values cannot be compared across countries or over time as all survey-years are standardized to have a zero mean.

estimating a principal component model using a set of asset variables that is commonly available in each survey, and using the factor scores to generate the measure of respondent's economic well-being. This results in a variable that though comparable across time and space, captures only ordinal differences in economic well-being. Thus, like nightlights, the variable has no substantively meaningful metric.

The most important difference between the Yeh et al approach and the approach here is of course that $\widehat{y^*_{ict}}$ is expressed as consumption per capita in 2011 PPP dollars. But there are other differences worth underscoring as well. First, the pooled PCA approach in Yeh et al requires a common set of assets in every survey. This limits the nature of the training data: since all surveys must contain the same set of asset variables, surveys must be dropped when they do not include the requisite asset variables. One way to limit the problem is to use a relatively small number of asset variables in constructing the measure of well-being. But this strategy limits the variability in the measure. By allowing the nature of the asset variables used in constructing $\widehat{y^*_{ict}}$ to vary across surveys, the approach here avoids both of these limitations. Second, the pooled PCA achieves comparability by assuming that the relationship between assets and economic well-being is the same across countries and over time. This is a strong assumption. We might expect, for example, that the relationship between owning a bicycle (or radio or computer or cell phone) and economic well-being to vary across countries or over time. The approach here avoids this strong assumption. Instead, it achieves comparability through the use of macro data on consumption and inequality. While this approach may have the problems discussed in section 3.2, the fact that $\widehat{y^*_{ict}}$ is denominated in dollars – and that it can be used to derive poverty rates – opens avenues for evaluating the measure that are unavailable when using measures lacking an interpretable metric.

B.3. From individuals to clusters. The final step is to average $\widehat{y^*_{ict}}$ to the enumeration area level, also called "cluster," which is roughly equivalent to villages in rural areas or neighborhoods in urban areas, as this is the level at which geo-coordinates are available in the public survey data. Since $\widehat{y^*_{ict}}$ is expressed in logs, we exponentiate it to obtain consumption per capita, compute the cluster-level average and then take the log of the average to obtain $\widehat{y^*_{rct}}$,

the log of cluster mean consumption in 2011 PPP dollars for cluster r, c, t.[4] For each cluster, DHS publishes the latitude and longitude of the cluster's "centroid." For privacy reasons, DHS randomly jiggers the published location of this centroid by up to 5 km from the true centroid. We therefore assign each DHS cluster to a 10x10 square kilometer pixel that has the centroid reported by DHS (thereby ensuring that the square encompasses the true centroid).[5] We denote this geo-located variable as $\widehat{y^*_{rct}}$, where r denotes enumeration area (cluster). The training data we use in the prediction models include 34,484 clusters.

[4]We drop clusters with invalid GPS coordinates. To limit measurement error due to small numbers of households in a cluster, we also drop clusters with less than 16 households (which results in dropping 2% of the clusters). On average, there are 26 respondents per cluster.

[5]Yeh et al (2020) estimate that the jiggering by DHS degrades model performance, reducing the R^2 by about 0.07.

C. Table describing predictors and their sources

Predictor	Definition	Source
nl_uncal_blur_10	DMSP nightlights (pre-2013) See Elvidege et al (1999) and Hsu et al (2015)	https://eogdata.mines.edu/products/dmsp/#docs
nl_uncal_deblur_10	DMSP deblurred nightlights (pre-2013) See Elvidege et al (1999) and Hsu et al (2015)	https://eogdata.mines.edu/products/dmsp/#docs
nl_viirs_10	VIIRS nightlight measure (post-2012) See Elvidge et al (2013) and Elvidge et al (2017)	https://eogdata.mines.edu/products/vnl/
nls_mean_yeh_c_b	Three-year moving average of NL, DSMP and VIIRS	Following Yeh et al (2020), this variable is created by standardizing nl_uncal_blur_10 and nl_viirs_10 and then taking the cell average of the standardized variable over three-year periods.
d_highway	Distance in meters from cell centroid to nearest highway. See Meijer et al (2018)	https://www.globio.info/download-grip-dataset
d_capital	Distance in meters from cell centroid to national capital	https://hub.arcgis.com/datasets/esri::world-cities/about
d_catholic	Distance in meters from cell centroid to nearest catholic mission	Cagé and Rueda (2020)
d_coast	Distance in meters from cell centroid to nearest coast	https://www.naturalearthdata.com/downloads/10m-physical-vectors/
d_diamonds	Distance in meters from cell centroid to nearest diamond deposit. See Gilmore et al (2005).	https://www.prio.org/data/10

Table continues on next page.

Table describing predictors and their sources, continued

Predictor	Definition	Source
d_harbor	Distance in meters from cell centroid to nearest harbor	http://msi.nga.mil/NGAPortal
d_lakes	Distance in meters from cell centroid to nearest lake	https://www.worldwildlife.org/pages/global-lakes-and-wetlands-database
d_missions	Distance in meters from cell centroid to nearest Christian mission	Nunn (2010)
d_offshoreoil	Distance in meters from cell centroid to nearest off-shore oil and gas deposit. See Lujala et al (2007).	https://www.prio.org/data/11
d_onshoreoil	Distance in meters from cell centroid to nearest on-shore oil and gas deposit. See Lujala et al (2007).	https://www.prio.org/data/11
d_protestant	Distance in meters from cell centroid to nearest protestant mission	Cagé and Rueda (2016)
d_rivers	Distance in meters from cell centroid to nearest river. See Lehner and Döll (2004)	https://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-rivers-lake-centerlines/
d_protected	Distance in meters from cell centroid to nearest protected area	https://www.protectedplanet.net/en/thematic-areas/wdpa?tab=WDPA
remoteness	Predicted score from a PCA that includes d_capital, d_catholic, d_coast, d_diamonds, d_harbor, d_missions, d_offshoreoil, d_onshoreoil, d_protestant, d_highway	
v_temp_avg_10	Average temperature in cell between 1960 and 1990 multiplied by 10. See Harris et al (2020).	https://crudata.uea.ac.uk/cru/data/hrg/
v_malariapf_10	Average prevalence of Malaria Plasmodium falciparum in cell	Weiss et al (2019)

Table continues on next page.

Table describing predictors and their sources, continued

Predictor	Definition	Source
v_rain_10	Average rain in cell in year. See Harris et al (2020).	https://crudata.uea.ac.uk/cru/data/hrg/
v_temp_10	Average temperature in cell in year multiplied by 10. See Harris et al (2020).	https://crudata.uea.ac.uk/cru/data/hrg/
disease	Predicted score from a PCA that includes v_malariapf_10, v_rain_10, v_temp_avg_10	–
v_elevation_10	Average elevation of cell, meters	Berry and Benveniste (2019)
v_calories_10	The average potential yields within each cell attainable given the set of crops	Galor and Özak (2015)
v_rugged_10	Terrain Ruggedness Index quantifying topographic heterogeneity in wildlife habitats providing concealment for preys and lookout posts	Nunn and Puga (2012)
geography	Predicted score from a PCA that includes v_calories_10, v_rugged_10, v_elevation_10	–
v_population_10	Cell population. See Doxsey-Whitfield (2015)	https://sedac.ciesin.columbia.edu/data/set/gpw-v4-population-count-rev11/data-download
v_co2_10	Sum of co2 emissions in each pixel within a square. See EDGARv7.0	https://edgar.jrc.ec.europa.eu/dataset_ghg70
latitude	Latitude of cell centroid	
longitude	Longitude of cell centroid	

Table continues on next page.

Table describing predictors and their sources, continued

Predictor	Definition	Source
eco1	Proportion of cell containing grassland	Olson and Dinerstein (2002)
eco3	Proportion of cell containing grasses and shrubs	Olson and Dinerstein (2002)
eco4	Proportion of cell containing a desert/xeric biome	Olson and Dinerstein (2002)

D. Additional information about the predictions models and their accuracy.

This section provides information about parameter tuning (section D.1) and about the variables with the largest importance in the prediction models (section D.2). It also presents additional information about prediction accuracy, including an analysis related to the poor performance of the models including only nightlights as predictors (section D.3). Finally, it compares the prediction results from our RF models with those of Yeh et al. (2020) (section D.5).

D.1. Parameter tuning. The main hyperparameters of the RF models are the number of individual trees (NTREES), the maximum number of predictors that are included in each tree (NVARS), maximum tree depth (DEPTH), and the minimum proportion of the variance at a node in order for splitting to be performed (VAR). To tune the different models, we consider a grid of values for each of the parameters. For each of the different values in the grid we estimate the random forest models using half of the sample; we then evaluate performance in the unseen data. Using this process, we identify the hyperparameter values leading to the lowest MSE for each model. Table D.1 presents the resulting hyperparameters employed in the three models.

	Preferred Hyperparameters Values				
	NTREES	NVARS	DEPTH	VAR	MIN OBS PER LEAF
MODEL 1	180	1	25	.0001	7
MODEL 2	180	8	35	.0001	3
MODEL 3	180	6	35	.0005	1

Table D.1. HYPERPARAMETER VALUES GENERATING THE LOWEST MSE IN THE THREE RF MODELS.

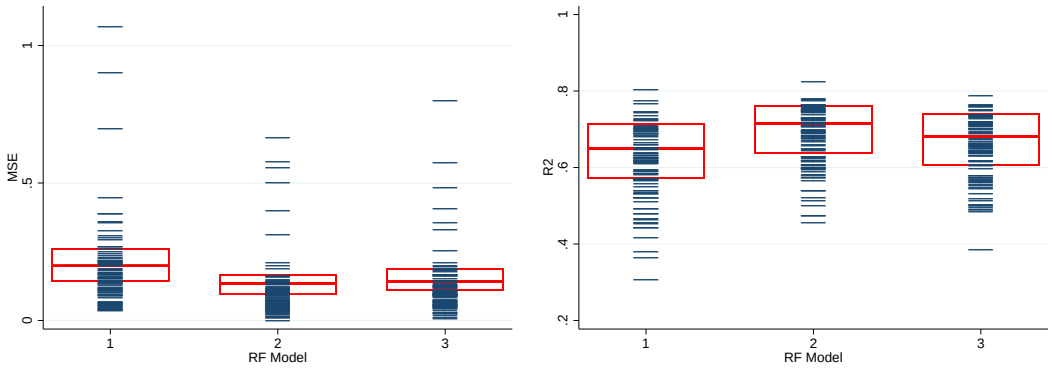
D.2. Variable importance. Table D.2 shows the most informative predictors in models RF-2 and RF-3. The most important variable in both models is whether a cell is located in the desert. Nightlights variables are important in RF-2, and CO2 emissions, population density, the disease environment, and variables related to a cell's location are important in both models.

D.3. Further evaluation of prediction accuracy. Figure D.1 displays the MSE and the R^2 from RF-1 through RF-3 using all 85 out-of-sample sets of forecasts. Both panels show that prediction accuracy varies across surveys. But this is especially true for RF-1. Focusing on panel (a), RF-1 generates a very large MSE for some surveys and substantial dispersion of

	Relative Variable Importance	
Ranking	RF-2	RF-3
1	Desert ecosystem (1)	Desert ecosystem (1)
2	NLs (3 yr mean) (.34)	CO_2 (.25)
3	CO_2 (.27)	Population Density (.16)
4	Population Density (.16)	Latitude (.08)
5	Latitude (.14)	Grassland ecosystem (.08)
6	Grassland ecosystem (.11)	Longitude (.07)
7	NL(VIIRS) (.10)	Remoteness (.07)
8	Longitude (.09)	Disease (.06)
9	Disease (.09)	Malaria Incidence (.05)
10	NL(DSMP, blur) (.09)	Grassland ecosystem (.05)

Table D.2. VARIABLE IMPORTANCE. This table provides the 10 most important predictors for models RF-2 and RF-3, together with their relative importance. Importance is relative to the most informative one (whose importance is normalized to 1).

the MSE across the surveys. In RF-2 and RF-3, by contrast, there are a half-dozen surveys that have especially poor performance (though much better than the MSE in RF-1), and the remaining MSEs are concentrated in quite low values, especially in our main model, RF-2.



(a) MSE, RF models **(b)** R^2, RF models

Figure D.1. OUT-OF-SAMPLE PREDICTION ACCURACY. This figure provides the MSE and R^2 for the 85 out-of-sample sets of predictions, corresponding to each of the surveys in our sample. Box and Whisker plots are displayed in red.

D.4. Understanding the poor performance of NL-only models. To understand the poor performance of the nightlights model, Figure D.2 depicts the squared correlation between the out of sample predictions and the training data (R^2) from RF-1 (NL only) and RF-2 (the full model) when increasing the number of deciles of data, X, used in estimation. For X=2, for instance, the graph depicts the value of the R^2 obtained when only the first two deciles

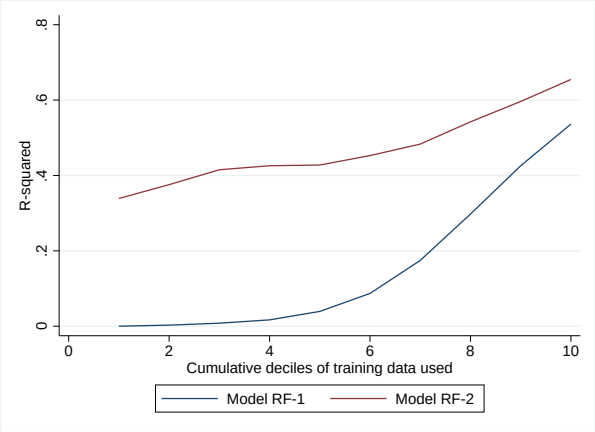


Figure D.2. PERFORMANCE FOR INCREASING SHARES OF DATA USED IN ESTIMATION. The figure plots the R^2s from models estimated on the X smallest deciles of the training data. E.g., if X=2, estimation is carried out on the first 2 deciles of the data.

of the training variable are used.[6] The graph shows that RF-1 forecast accuracy is basically zero when as much as the first six deciles are used for estimation, and it remains quite low until 90% of the observations are employed. The graph therefore confirms that nightlights alone have no power to predict variation in economic well-being for almost 60% of the data. What nightlights make possible is to distinguish the 90% of poorest clusters from the 10% of richest ones. This is unsurprising given the vast areas of populated darkness described in the Introduction. By contrast, the performance of RF-3 is is stronger across the deciles, with an R^2 around .5 for these first five deciles, which then grows by over 20% when the remaining deciles are included in the estimation sample. This graph therefore highlights that in comparison to the NL-only models, the models with a richer set of predictors not only have power to distinguish the poor from the rich, but also to distinguish the poor from the very poor.

D.5. Comparison with previous benchmarks: Yeh et al. (2020). Yeh et al (2020) present an innovative method based on combining nightlights and daytime imagery to predict an asset wealth index, which is computed using 43 DHS surveys across almost 20,000 African clusters. They train a convolutional neural network (CNN) to predict the cluster-specific

[6]Out-of-sample predictions are obtained as follows: (1) the clusters in the training data are divided into ten "decile" data sets, each including the clusters in the decile and all lower deciles; (2) for each decile data set, RF-1 and RF-2 are estimated, but omitting the data from a held out survey; (3) out-of-sample predictions are obtained for the held-out clusters in each decile data set; (4) the R^2 for each decile data set is obtained using the out-of-sample predictions.

measure of wealth using temporally and spatially matched multispectral daytime imagery and nightlights as inputs. Based on prediction performance on held out locations, they show that (1) the median squared correlation coefficient between their training variable and the (out-of-sample) predictions is 70% in their best models, and (2) a simple K-nearest neighbor (KNN) model whose only predictor is nightlights has similar performance to a CNN with both nightlights and (the much heavier) daylight imagery.

Our framework is similar to that of Yeh et al. (2020), but there are several differences: (1) they use a different training variable (a unitless index of asset wealth, computed as the first principal component in a sample that pools all DHS surveys), (2)) their sample is much smaller (43 surveys versus 85, in our case), and (3) their country composition is different. Consequently, the results presented in section 4.3 are not directly comparable to the results in Yeh et al. To investigate the relative performance of both approaches, in this section we use random forest models with a large number of variables to predict the training variable from the Yeh et al data set. This allows us to attribute any performance differences to the different methods and predictors we employ.

	MSE and R^2 (median value))	
	MSE	R^2
Yeh, KNN	.191	.691
Yeh, CNN	.179	.687
Random forest, with rich set of predictors	.168	.724

Table D.3. Comparing predictions results from Yeh et al (2020) with those from random forest models using the Yeh et al training data. Table provides model results when using the training variable from the Yeh et al data set. TheYeh, KNN and Yeh, CNN results are obtained from the Yeh et al replication materials.

For the results from Yeh et al, we focus on (a) their model with best performance, a CNN that uses both NLs and daylight imagery as inputs, and (b) their KNN model that includes only NLs (as the latter is much less computationally intensive, and Yeh et al. conclude that its performance is very similar to that of more complex models). We compare performance to a random forest model that includes a rich set of cell-level predictors, as we do in RF-2. All results are based on held-out locations, as described in the main text. Table D.3 presents the results.

Two conclusions stand out. First, when the preferred MSE metric is employed, the performance of the CNN model is significantly better than that of the simpler KNN model (an improvement of 6% in the median MSE). Second, the random forest model with a rich set of predictors outperform those in Yeh et al. (2020), producing a MSE that is 12% smaller than that of the KNN model and 6% smaller than that of the CNN model. Importantly, these improved results are achieved at a much lower computational cost. The use of daytime imagery together with the CNN algorithm is a very expensive computational approach. This complexity and computational expense make it extremely challenging to scale up the predictions to compute maps for the whole of Africa over time. By contrast, the random forest models with a large number of predictors can easily be run in STATA on a personal computer.

E. Estimating cell-level poverty rates

This section provides further details about the nonparametric approach we propose for estimating cell poverty rates. We must first classify clusters into K groups, and then estimate the poverty rate of clusters belonging to group k, for $k = 1 \ldots K$. We do this in three steps:

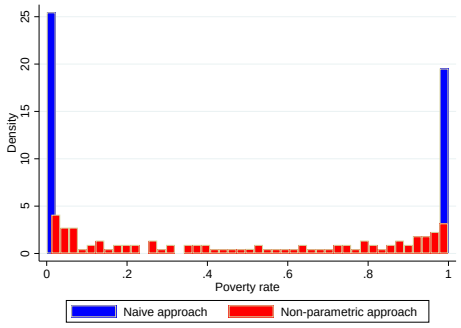
(1) We assume $K = 100$. To allocate clusters to groups we could use clustering methods, such as k-means. However, we will assume that all clusters in group k have nearly identical average consumption, making it natural to use consumption to assign clusters to groups. Therefore, groups are defined as percentiles of the distribution of $\widehat{y^*_{rct}}$. We use the percentile values of each group to identify the cut-points dividing a group from its adjacent groups. For example, the 75th group includes all clusters that have a value of $\widehat{y^*_{rct}}$ between 7.194 and 7.227. Like group 75, the range of cluster mean consumption in each group is very narrow, with an average range that is 0.031.

(2) We assign each of the roughly 920,000 respondents in the DHS surveys to the group associated with the respondent's cluster mean income. Since the 75th group includes all clusters that have a value of $\widehat{y^*_{rct}}$ between 7.194 and 7.227, if a DHS survey respondent resides in a cluster, for example, with $\widehat{y^*_{rct}}$= 7.21, the respondent would be placed in the 75th group, along with all other respondents across all surveys who reside in a cluster

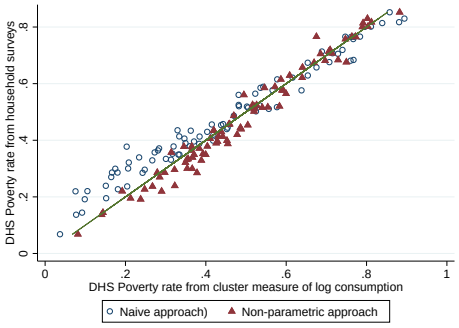
with a value of $\widehat{y^*_{rct}}$ between 7.194 and 7.227. This particular group has 8,790 individual DHS respondents, which is a typical number of survey respondents in each group.

(3) We use all the DHS respondents in a group (such as the 8,790 in group 75) to calculate the poverty line for the group. Our focus will be on the \$1.90 per day used with 2011 PPP dollars. Using the 8,790 estimates of $\widehat{y^*_{rct}}$ in group 75, we find that 19.1% of the households are below the poverty line of \$1.90 per day using 2011 PPP dollars. Thus, the poverty rate for group 75 is 19.1% and all clusters assigned to this group are assigned this same poverty rate.

Panel (a) in Figure E.1 shows the histogram of cluster-level poverty rates based on $\widehat{y^*_{rct}}$. The blue histogram uses the naive approach, and the stark binary nature of the distribution is what motivates the nonparametric method. The red histogram uses the nonparametric approach to generate the poverty rates, and though there is still slightly more mass at the tails of the distribution, there is a relatively even distribution across the range of poverty rates.



(a) Naive versus Non-parametric: Histogram



(b) Naive versus Non-parametric: Binned scatterplots

Figure E.1. The distribution of cluster-level poverty rates using naive and non-parametric approaches. Panel (a) depicts the histogram of poverty rates in DHS clusters using the naive and non-parametric approaches. The poverty line is \$1.90 a day using 2011 PPP dollars. Panel (b) presents scatterplots of log consumption versus (transformed) asset-based indices, along with the 45-degree line.

By aggregating these rates to the country level using using DHS weights, we can gain further insights into how the two approaches differ. We have calculated the national level poverty rates based on the individual-level data, $\widehat{y^*_{ict}}$. Panel (b) in Figure E.1 compares these poverty rates (the y-axis) with the poverty rates based on applying the naive and non-parametric approaches to the cluster-level data. When aggregated to the national level, both

approaches use the cluster-level data to produce national poverty rates that are close to those from the individual-level data. But the naive approach tends to underestimate poverty at the higher end of the poverty distribution, and especially, to overestimate poverty at the lower end of the poverty distribution. The same patterns do not exist for the non-parametric estimates. As noted, the goal of the non-parametric approach is to produce fine-grained estimates of poverty rates for cell-level analyses. The evidence here suggests that by doing so, we also replicate national poverty rates more accurately than does the naive approach.

F. Additional results from re-estimating models in MP13 and MP14 using SED data

This section presents additional results related to re-analysis of MP13 and MP14 models.

Additional MP14 models. Tables F.1 to F.3 present results from models in MP14. Each table states the models from MP14 that are being re-estimated, and the reader is referred the tables in MP14 for model specifics. Together, Tables F.1 to F.3 show that there is no robust relationship between national institutions and nightlights, as in MP14. By contrast, the results for RULE OF LAW are positive and precisely estimated in every model, and the results for CONTROL OF CORRUPTION are precisely estimated in all models except the spatial regression discontinuity models in MP 14 Table VI.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Panel A: Dep. variable is NL				
RULE OF LAW	0.0994***	0.0168	0.1031**	0.0159				
	(0.0384)	(0.0181)	(0.0407)	(0.0162)				
CONTROL OF CORRUPTION					0.1292***	0.0269	0.1346***	0.0197
					(0.0471)	(0.0259)	(0.0492)	(0.0210)
Adj. R-squared	0.134	0.319	0.137	0.345	0.144	0.319	0.149	0.345
N	21289	21289	13408	13408	21289	21289	13408	13408
				Panel B: Dep. variable is consumption, RF-2				
RULE OF LAW	0.4474***	0.2382***	0.4365***	0.2139***				
	(0.1677)	(0.0800)	(0.1590)	(0.0768)				
CONTROL OF CORRUPTION					0.5823***	0.2794***	0.5650***	0.2367**
					(0.1598)	(0.1017)	(0.1530)	(0.0982)
Adj. R-squared	0.227	0.773	0.230	0.784	0.295	0.770	0.296	0.780
N	20441	20441	12869	12869	20441	20441	12869	12869
Ethnicity fixed effects	No	Yes	No	Yes	No	Yes	No	Yes
Population density and area	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Location controls.	No	No	Yes	Yes	No	No	Yes	Yes
Geographic controls	No	No	Yes	Yes	No	No	Yes	Yes

Table F.1. RE-ESTIMATING MP14 TABLE V, PANEL B. This table re-estimates MP14 Table 5, panel B.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: Dep. variable is NL												
RULE OF LAW (HIGH)	0.0166	0.0058	0.0123	0.0050	-0.0010	0.0122						
	(0.0138)	(0.0135)	(0.0165)	(0.0185)	(0.0239)	(0.0271)						
CONTROL OF CORRUPTION (HIGH)							0.0038	-0.0086	0.0116	0.0123	-0.0030	0.0079
							(0.0154)	(0.0137)	(0.0166)	(0.0176)	(0.0225)	(0.0278)
Adj. R-squared	0.342	0.342	0.320	0.320	0.347	0.347	0.342	0.343	0.320	0.320	0.347	0.347
N	40209	40209	21289	21289	13408	13408	40209	40209	21289	21289	13408	13408
Panel B: Dep. variable is consumption, RF-2												
RULE OF LAW (HIGH)	0.1268***	0.1292***	0.0920**	0.1069***	0.1037***	0.1155**						
	(0.0415)	(0.0384)	(0.0389)	(0.0384)	(0.0402)	(0.0449)						
CONTROL OF CORRUPTION (HIGH)							0.0392	0.0338	0.0145	0.0316	0.0166	0.0232
							(0.0548)	(0.0531)	(0.0483)	(0.0479)	(0.0503)	(0.0596)
Adj. R-squared	0.798	0.798	0.775	0.775	0.786	0.786	0.786	0.786	0.760	0.760	0.771	0.771
N	38438	38438	20441	20441	12869	12869	38438	38438	20441	20441	12869	12869
Ethnicity fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pixel area and pop. dens.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table F.2. RE-ESTIMATING MP14 TABLE VI. This table re-estimates MP14 Table VI, which estimates border regression discontinuity models.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: Dep. variable is NL												
RULE OF LAW (HIGH)	0.0159	0.0116	-0.0027	0.0134	0.0049	-0.0207						
	(0.0186)	(0.0199)	(0.0212)	(0.0311)	(0.0376)	(0.0330)						
CONTROL OF CORRUPTION (HIGH)							0.0090	0.0133	-0.0030	0.0410	0.0367	0.0024
							(0.0205)	(0.0214)	(0.0197)	(0.0313)	(0.0313)	(0.0230)
Adj. R-squared	0.398	0.400	0.400	0.458	0.461	0.462	0.399	0.399	0.400	0.457	0.460	0.461
N	19349	19349	19349	10031	10031	10031	19814	19814	19814	10353	10353	10353
Panel B: Dep. variable is consumption, RF-2												
RULE OF LAW (HIGH)	0.1941**	0.1440***	0.1287**	0.2457	0.2629***	0.2007***						
	(0.0979)	(0.0557)	(0.0513)	(.)	(0.0890)	(0.0729)						
CONTROL OF CORRUPTION (HIGH)							0.2506**	0.1622**	0.1449**	0.3010***	0.2185***	0.1869***
							(0.0994)	(0.0652)	(0.0634)	(0.1127)	(0.0743)	(0.0694)
Adj. R-squared	0.726	0.728	0.728	0.739	0.745	0.748	0.779	0.782	0.783	0.796	0.803	0.804
N	17796	17796	17796	8478	8478	8478	18261	18261	18261	9241	9241	9241
Ethnicity fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Pixel area and pop. dens.	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Table F.3. RE-ESTIMATING MP14 TABLE VII. This table re-estimates MP14 Table VII, which estimates border regression discontinuity models when institutional differences across borders are large.

Additional MP13 models. Tables F.4 to F.7 present results from models in MP13 to demonstrate there is no robust relationship between JURISDICTIONAL HIERARCHY when SED consumption is used as the outcome variable. Each table describes the model being re-estimated and the reader can find details about model specifics in MP13.

	(1)	(2)	(3)	(4)	(5)
	Panel A: Dep. variable is continuous lights				
JURISDICTIONAL HIERARCHY	0.1343	0.1529*	0.1029**	0.1176***	0.0793***
	(0.0969)	(0.0886)	(0.0450)	(0.0454)	(0.0305)
Adjusted R-squared	0.01	0.19	0.28	0.32	0.33
Observations	61359	61359	61359	61015	61015

Table F.4. RE-ESTIMATING MP13, TABLE V PANEL A. This table re-estimates MP13 Table V, Panel A using the MP13 measure of log of nightlights as outcome variable.

	(1)	(2)	(3)	(4)	(5)
	Panel A: Dep. variable is NL				
Petty Chiefdoms	0.0120	0.0459	0.0290	0.0194	0.0113
	(0.0231)	(0.0346)	(0.0225)	(0.0182)	(0.0137)
Paramount Chiefdoms	0.0538	0.0843*	0.0602*	0.0642**	0.0464**
	(0.0340)	(0.0507)	(0.0311)	(0.0306)	(0.0187)
Pre-Colonial States	0.0853	0.0990*	0.0638**	0.0625***	0.0383**
	(0.0647)	(0.0511)	(0.0252)	(0.0223)	(0.0176)
Adj. R-squared	0.008	0.182	0.268	0.288	0.294
N	61359	61359	61359	61015	61015
	Panel B: Dep. variable is consumption from RF-2				
Petty Chiefdoms	-0.1631*	0.0487	0.0333	0.0012	0.0070
	(0.0949)	(0.0463)	(0.0439)	(0.0397)	(0.0427)
Paramount Chiefdoms	-0.1040	-0.0147	-0.0368	-0.0221	-0.0096
	(0.1530)	(0.0691)	(0.0626)	(0.0337)	(0.0331)
Pre-Colonial States	-0.0434	0.0076	-0.0246	-0.0425	-0.0505
	(0.1842)	(0.0629)	(0.0696)	(0.0567)	(0.0624)
Adj. R-squared	0.007	0.778	0.793	0.832	0.836
N	61359	61359	61359	61015	61015
Country Fixed effects	No	Yes	Yes	Yes	Yes
Population Density	No	No	Yes	Yes	Yes
Controls at the Pixel level	No	No	No	Yes	Yes
Controls at the Ethnic-Country level	No	No	No	No	Yes
Observations	61359	61359	61359	61015	61015

Table F.5. Re-estimating MP13, Table V panel B. This table re-estimates MP13 Table V, Panel B using the transformed consumption variable from RF-2.

Should we get rid of the NL rows in the tables that follow?

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: Dep. variable is NL									
JURISDICTIONAL HIERARCHY	0.0124	0.0077	0.0043	0.0118	0.0084	0.0044	0.0161	0.0115	0.0046
	(0.0094)	(0.0050)	(0.0045)	(0.0115)	(0.0060)	(0.0053)	(0.0168)	(0.0085)	(0.0051)
Adj. R-squared	0.259	0.314	0.323	0.218	0.287	0.295	0.278	0.332	0.343
N	72545	72545	72297	30008	30008	29915	12851	12851	12812
Panel B: Dep. variable is consumption from RF-2									
JURISDICTIONAL HIERARCHY	-0.0155	-0.0202	-0.0188	-0.0215	-0.0242	-0.0261	-0.0024	-0.0055	0.0061
	(0.0224)	(0.0247)	(0.0225)	(0.0277)	(0.0305)	(0.0275)	(0.0255)	(0.0283)	(0.0195)
Adj. R-squared	0.855	0.867	0.880	0.869	0.876	0.894	0.904	0.908	0.916
N	72545	72545	72297	30008	30008	29915	12851	12851	12812
Adjacent-Ethnic-Groups Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Population Density	No	Yes	Yes	No	Yes	Yes	No	Yes	Yes
Controls at the Pixel level	No	No	Yes	No	No	Yes	No	No	Yes

Table F.6. RE-ESTIMATING MP13, TABLE VII. The dependent variable is consumption from RF-2.

	Dependent Variable: Consumption		
	< 100 km of ethnic border (1)	< 150 km of ethnic border (2)	< 200 km of ethnic border (3)
Panel A: Pre-Colonial Ethnic Institutions and Regional Development Within Contiguous Ethnic Homelands in the Same Country Pixel-Level Analysis in Areas Close to the Ethnic Border			
Panel 1: Border Thickness: Total 50 km (25 km from each side of the ethnic boundary)			
JURISDICTIONAL HIERARCHY	0.0168	0.0094	0.0105
	(0.0173)	(0.0158)	(0.0158)
Adj. R-squared	0.908	0.898	0.897
N	6237	9476	11920
Panel 2: Border Thickness: Total 100 km (50 km from each side of the ethnic boundary)			
JURISDICTIONAL HIERARCHY	0.0117	0.0032	0.0066
	(0.0178)	(0.0161)	(0.0166)
Adj. R-squared	0.906	0.896	0.896
N	4053	7292	9736
Panel B: Pre-Colonial Ethnic Institutions and Regional Development Within Contiguous Ethnic Homelands in the Same Country Pixel-Level Analysis in Areas Close to the "Thick" Ethnic Border Border Controlling for a Fourth-order RD-Type Polynomial in Distance to the Ethnic Border			
Panel 1: Border Thickness—Total 50 km (25 km from each side of the ethnic boundary)			
JURISDICTIONAL HIERARCHY	0.0104	0.0145	0.0205
	(0.0272)	(0.0248)	(0.0231)
Adj. R-squared	0.909	0.899	0.898
N	6237	9476	11920
Panel 2: Border Thickness—Total 100 km (50 km from each side of the ethnic boundary)			
JURISDICTIONAL HIERARCHY	0.0128	0.0174	0.0220
	(0.0397)	(0.0337)	(0.0302)
Adj. R-squared	0.906	0.896	0.896
N	4053	7292	9736
RD-Type Polynomial	Yes	Yes	Yes
Adjacent-Ethnic-Groups Fixed Effects	Yes	Yes	Yes
Population Density	Yes	Yes	Yes
Controls at the Pixel level	Yes	Yes	Yes

Table F.7. RE-ESTIMATING MP13, TABLE VIII USING CONSUMPTION. This table re-estimates the first three models (which use all observations) from MP13 Table VIII. The dependent variable is consumption from RF-2.

G. References

Battistin, Erich, Blundell, Richard, and Lewbel, Arthur (2009), "Why Is Consumption More Log Normal than Income? Gibrat's Law Revisited," *Journal of Political Economy*, 117 , 1140–1154.

Berry, P. A. M., Smith, R., and Benveniste, J., (2010), "ACE2: The New Global Digital Elevation Model" In Mertikas, S. (eds) Gravity, Geoid and Earth Observation, International Association of Geodesy Symposia, Berlin, Heidelberg, Springer.

Cagé, Julia, and Valeria Rueda,(2016) "The Long-Term Effects of thePrinting Press in sub-Saharan Africa," *American Economic Journal: Applied Economics* 8, 69–99.

Cage, Julia, and Valeria Rueda, (2020), "Sex and the Mission: The Conflicting Effects of Early Christian Investments on the HIV Epidemic in Sub-Saharan Africa," *Journal of Demographic Economics* , Volume 86 ,(3). 213 - 257.

Chi, G, Fang, S, and Blumenstock, JE (2022). "Microestimates of Wealth and Poverty for all Low- and Middle-Income Countries," *Proceedings of the National Academy of Sciences*, 119(3), 1-11.

Doxsey-Whitfield, Erin, Kytt MacManus, Susana B. Adamo, Linda Pistolesi, John Squires, Olena Borkovska and Sandra R. Baptista (2015), "Taking Advantage of the Improved Availability of Census Data: A First Look at the Gridded Population of the World, Version 4," *Papers in Applied Geography*, 1:3, 226-234, DOI: 10.1080/23754931.2015.1014272.

EDGAR (Emissions Database for Global Atmospheric Research) Community GHG Database (a collaboration between the European Commission, Joint Research Centre (JRC), the International Energy Agency (IEA), and comprising IEA-EDGAR CO2, EDGAR CH4, EDGAR N2O, EDGAR F-GASES version 7.0, (2022) European Commission, JRC (Datasets).

Elvidge, Christopher D., Baugh, Kimberly E., Dietz, John B., Bland, Theodore, Sutton, Paul C., and Kroehl, Herbert W., (1999), "Radiance Calibration of DMSP-OLS Low-Light Imaging Data of Human Settlements," *Remote Sensing of Environment*, 68, 77–88.

Elvidge, Christopher D, K. Baugh, M. Zhizhin, F. C. Hsu, and T. Ghosh, (2017), "VIIRS night-time lights," *International Journal of Remote Sensing*, vol. 38, pp. 5860–5879.

Elvidge, Christopher D., Baugh, Kimberly E., Zhizhin, Mikhail, and Hsu, Feng-Chi, (2013), "Why VIIRS data are superior to DMSP for mapping nighttime lights," *Proceedings of the Asia-Pacific Advanced Network*, 35 (2013), 62.

Galor, Oded, and Zak, Mer, (2015) "Land Productivity and Economic Development: Caloric Suitability vs. Agricultural Suitability," *Agricultural Suitability*, July 12, 2015.

Gibson, J., O. Susan, and G.Boe-Gibson (2020), "Night Lights in Economics: Sources and Uses," *Journal of Economic Surveys* 34, 955–980.

Gilmore, Elisabeth, Gleditsch, Nils Petter, Lujala, Päivi, and Ketil Rød, Jan, (2005) "Conflict diamonds: A new dataset," *Conflict Management and Peace Science*, 22, 257–272.

Harris, Ian, Osborn, Timothy J., Jones, Phil, and Lister, David, (2020), "Version 4 of the CRU TS monthly high-resolution gridded multivariate climate dataset,// *Scientific Data*, 7, 109.

Hsu, F. C., Baugh, K. E., Ghosh, T., Zhizhin, M., Elvidge, C. D. (2015), "DMSP-OLS radiance calibrated nighttime lights time series with intercalibration," *Remote Sensing*, 7(2), 1855-1876.

Lehner, Bernhard, and Petra Döll, (2004), "Development and validation of a global database of lakes, reservoirs and wetlands," *Journal of hydrology*, 296, 1–22.

Lujala, Päivi; Jan Ketil Rød and Nadia Thieme, (2007), "Fighting over Oil: Introducing A New Dataset," *Conflict Management and Peace Science* 24(3), 239-256.

Meijer, Johan R, Huijbregts, Mark AJ, Schotten, Kees CGJ, and Schipper, Aafke M, (2018), "Global patterns of current and future road infrastructure," *Environmental Research Letters*, 13 (2018), 064006.

Michalopoulos, S., and E. Papaioannou (2013), "Pre-Colonial Ethnic Institutions and Contemporary African Development," *Econometrica* 81, 113–152.

Michalopoulos, S., and E. Papaioannou (2014), "National Institutions and Subnational Development in Africa," *The Quarterly Journal of Economics* 129, 151–213.

Nunn, Nathan, (2010), "Religious Conversion in Colonial Africa," *American Economic Review*, 100 (2010), 147–152.

Nunn, Nathan, and Puga, Diego, "Ruggedness: The Blessing of Bad Geography in Africa," *Review of Economics and Statistics* 94 (2012), 20–36.

Olson, David M, and Dinerstein, Eric, (2002), "The Global 200: Priority ecoregions for global conservation," *Annals of the Missouri Botanical Garden*," 199–224.

Weiss, Daniel J, Lucas, Tim C D, Nguyen, Michele, Nandi, Anita K, Bisanzio, Donal, Battle, Katherine E, Cameron, Ewan, Twohig, Katherine A, Pfeffer, Daniel A, Rozier, Jennifer A, Gibson, Harry S, Rao, Puja C, Casey, Daniel, Bertozzi-Villa, Amelia, Collins, Emma L, Dalrymple, Ursula, Gray, Naomi, Harris, Joseph R, Howes, Rosalind E, Kang, Sun Yun, Keddie, Suzanne H, May, Daniel, Rumisha, Susan, Thorn, Michael P, Barber, Ryan, Fullman, Nancy, Huynh, Chantal K, Kulikoff, Xie, Kutz, Michael J, Lopez, Alan D, Mokdad, Ali H, Naghavi, Mohsen, Nguyen, Grant, Shackelford, Katya Anne, Vos, Theo, Wang, Haidong, Smith, David

L, Lim, Stephen S, Murray, Christopher J L, Bhatt, Samir, Hay, Simon I, and Gething, Peter W, (2019), "Mapping the global prevalence, incidence, and mortality of Plasmodium falciparum, 2000–17: a spatial and temporal modelling study," *The Lancet*, 394 (2019), 322–331.

Yeh, C., A. Perez, A. Driscoll, G. Azzari, Z. Tang, D. Lobell, S. Ermon, and M. Burke (2020) "Using publicly available satellite imagery and deep learning to understand economic well-being in Africa," *Nature Communications*, 11, 2583.