Economic development in pixels [1]

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

February 2025

Abstract

We describe a novel and easily implementable methodology for generating estimates of per capita consumption in 10x10 km cells over time. The first step in this methodology is based on a mathematical framework for translating asset indices from surveys into measures of consumption. The framework allows us to develop a training variable for machine learning models, which we use in a second step to create estimates of consumption for 42 sub-Saharan African countries. These new data allow us to highlight a fundamental problem with using "nightlights" – a *de facto proxy* for local economic well-being – in statistical models. We do so by revisiting two prominent papers that examine the effect of institutions on economic development, both of which rely on nightlights. The conclusions from these papers are reversed when we substitute our consumption-based measure for nightlights, and we show that this reversal is due to the nonclassical measurement error that is endemic to nightlights in regions where large swaths of territory appear unlit, as is common in rural Africa. This error can introduce unpredictable biases – either attenuating or amplifying – into statistical models that use nightlights as a measure of spatial economic performance. In contrast to nightlights, our methodology makes it possible to remove non-classical measurement error from the consumption estimates, and it produces estimates denominated in a metric that is well-understood. The approach described here therefore provides a promising way forward in the study of a wide range of questions and policies that have so far relied on nightlight data to track economic progress.

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 significant limitations to using nightlights as a spatially fine-grained proxy for development. First, while nightlights are correlated with development, they do not directly measure it (as consumption or income do). Consequently, they lack a substantively interpretable metric, offering at best ordinal information about spatial economic activity. Results in papers analyzing nightlights therefore typically express estimated effects in terms of "change in log lumens" or "change in probability a pixel is lit," quantities that have quite limited economic interpretability. Second, even if scholars using nightlights take extra steps to improve substantive interpretation of estimated effects, these effects will often be biased in an unknown direction. This is true because nightlights, when used in spatially disaggregated models, suffer nonclassical measurement error. That nightlights contain measurement error has of course been widely recognized, including by Chen and Nordhaus (2011). But the fact that measurement error is non-classical poses a considerable challenge for research using nightlights as a proxy for spatially disaggregated development

[2]See Donaldson and Storeygard (2016), 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.

in regression models because estimates from such models will typically be biased in *ex ante* unpredictable ways, regardless of whether nightlights is used as dependent or independent variable. Although some authors have warned about certain difficulties posed by nonclassical measurement error in nightlights,[3] the scope of the problem has not to our knowledge been previously recognized, and applied researchers have not paid sufficient attention to these challenges, as shown by the explosion of papers using nightlights in applied research in recent years (see Figure A.1 in Appendix A).

Motivated by these limitations of nightlights and their prevalence in development research, this paper attempts to make three distinct contributions. First, we develop a novel and easily implementable approach for using machine learning to create estimates of consumption per capita in 10x10km squares, and we implement this approach in sub-Saharan Africa over time. Second, we describe a simple method for ridding the consumption estimates of non-classical measurement error so that the new data can be employed in regression analysis using standard techniques. Third, we use the new consumption estimates to highlight a specific source of non-classical measurement error in nightlights, and we then show how this measurement error can be linked to both amplification and attenuation bias in regression models. The analysis therefore suggests that the use of nightlights in spatially disaggregated regressions may often lead to fundamentally flawed conclusions. We illustrate the problem by revisiting two well known papers on institutions and economic development. The remainder of this Introduction describes these contributions in further detail.

Our approach to developing the spatially disaggregated estimates of consumption per capita involves three novel steps. First, motivated by the fact that in large parts of the world there is no high resolution data on consumption or income per capita, whereas data on individual asset ownership often exists, we develop a mathematical framework for using country-level data on the distribution of consumption to translate an individual asset index (derived from surveys) into an individual-level measure of log consumption. The framework is built on transparent assumptions that we can test.

[3]For instance, Gibson, Olivia and Boe-Gibson (2020) argue that problems such as overglow or top-coding generate mean-reverting measurement errors which leads to attenuation in coefficient estimates even when nightlights are used as dependent variable.

Asset indices are widely used directly as individual-level measures of economic well-being in studies of development (e.g., Young (2013), Acemoglu, Reed and Robinson (2014), and Lowes and Montero (2021)), and they have more recently been used to create training variables for machine learning models (e.g., Jean et. al. (2016), Yeh et al. (2020 and 2021), Chi et al. (2022) and Aiken et al. (2022)).[4] These indices are attractive in part because it is often difficult to measure individual-level income or consumption accurately in some developing parts of the world. But asset indices lack an interpretable metric, and they require the strong assumption that the economic value of particular assets is the same across time and place.[5] By translating asset indices into consumption dollars, we avoid both problems. Thus, this approach is not only valuable for the machine-learning application in this paper, but also for any research that uses asset indices to gain deeper insights into economic well-being.

The second step is to use machine learning to create spatially disaggregated estimates of consumption per capital in PPP dollars. We implement the previously introduced mathematical framework 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 country-level consumption and inequality.[6] Together, these data sources allow us to create the asset indices and translate them into an individual-level measure of consumption dollars. Since the data are geocoded, we can average the individual-level consumption values for respondents in the same location, creating a geocoded measure of consumption per capita that will be used as a training variable. We use multiple data sources to validate this new consumption measure and the assumptions employed to create it.

This new training variable is used in random forest models that use a wide variety of predictors – including nightlights, as well as a predictors related to a cell's remoteness, geography, disease environment, weather fluctuations, CO2 emissions, population, and characteristics of the ecosystem, among others – to predict consumption in the DHS cells. The models achieve high prediction accuracy, outperforming existing methods at much lower computational cost. This accuracy is robust when different sets of predictor variables are

[4]These papers also employ the DHS survey data that we use to implement our approach.

[5]See section 2.3 for additional discussion of using asset indices in machine learning.

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

utilized. The trained models make possible the creation of estimates of log consumption per capita across 10x10km cells in 42 sub-Saharan African countries over 15 years. Although there are over 4 million observations in the resulting data set, the approach is very computationally efficient, running quickly on a laptop computer. Thus, although sub-Saharan Africa is the focus here, because of its simplicity, robustness and computational efficiency, the approach we develop can easily be implemented in other contexts.[7]

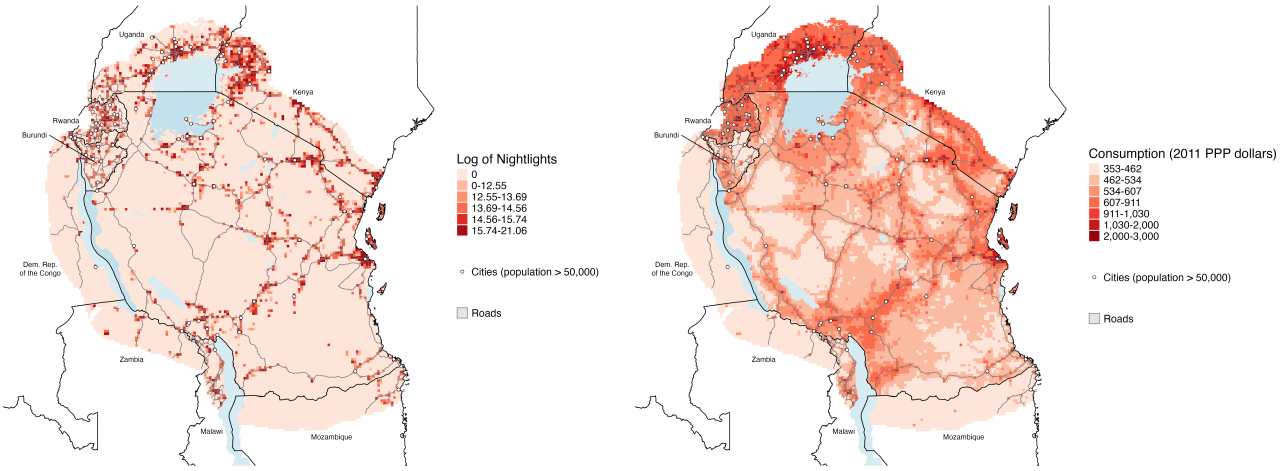
The third step in creating the new consumption data addresses the removal of non-classical measurement error from the random forest predictions. Such error is a common challenge when generating data via machine learning and is present in our measures of consumption. Consequently, relying on these estimates in regression models would lead to biased coefficient estimates. Solving this problem is an active area of research, and although a comprehensive treatment is beyond the scope of this paper, we propose a simple approach that uses the training variable to eliminate the non-classical measurement error in the newly generated data.[8] This approach, which constitutes our paper's second contribution, can be applied to other settings where the training data is representative of the target data.

Figure 1 illustrates how the new consumption estimates differ from nightlights. Panel (a) shows a map of nightlights in Tanzania in 2013 (a year when the more accurate VIIRS data is available).[9] The most striking feature of the map is the vast swaths of darkness: 89% of cells in Tanzania are dark. Panel (b) shows a map of consumption per capita using the data we generate. In addition to the fact that the cells are denominated in dollars, there are a number of clear differences from the nightlights map, including 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.

[7] See further discussion in section 3.3.

[8] See Battaglia et al. (2024) and the references therein for recent contributions on this problem.

[9] 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 estimates (from RF-2) developed in this paper.

These new consumption estimates make possible our third contribution, which is to highlight a significant limitation of using nightlights in spatially disaggregated analyses – one that to our knowledge has not received the attention it deserves. Specifically, we demonstrate that the well-known low sensitivity of satellite sensors results in vast areas where no light is recorded, a problem we show is acute in Africa. We argue that this issue is not merely one of data censoring but of misclassification, where dark areas are incorrectly assigned to have no economic activity. This missclassification introduces *a negative correlation between measurement error in nightlights and the true level of economic development*. The survey data we use to compute our measures of consumption show that considerable economic activity often exists in these "dark" areas, driving this negative correlation. The non-classical measurement error arising from this misclassification biases coefficient estimates in regression models that use nightlights in a spatially disaggregated fashion, either as a dependent or independent variable. Since this bias can manifest in either direction, it renders the interpretation of these coefficients inherently unreliable and potentially misleading. We would emphasize, however, that these concerns about this specific source of non-classical measurement error pertain specifically to the use of nightlights in spatially disaggregated analyses. When nightlights are employed at higher

levels of aggregation, such as at the country level, the concern we emphasize here does not necessarily apply.[10]

We illustrate the implications of this bias by revisiting two influential papers, Michalopoulos and Papaioannou (2013 and 2014), that use nightlights as a dependent variable to study the effect of institutions on economic development. Michalopoulos and Papaioannou (2013) use nightlights as a dependent variable to show that centralized ethnic institutions positively affect economic development. 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 economic development, though good institutions possibly encourage development in locations that are very close to a country's capital. We use the data and statistical models from these two papers but substitute for nightlights the new estimates of consumption per capita, adjusted to eliminate non-classical measurement error.

Strikingly, the results are reversed compared to those reported in the original studies. We find no effect of centralized ethnic institutions on development, and a strong positive effect of national institutions on development, regardless of a cell's location. Why the change in results? 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 (explaining how Michalopoulos and Papaioannou (2013) can obtain significant results when the true effect of ethnic institutions is likely zero). We also 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 (explaining how Michalopoulos and Papaioannou (2014) can obtain null results when in fact national institutions likely have a positive effect on development).

Not only are the results fundamentally different when we use the measure of consumption, they are much easier to interpret. For example, substantive interpretation in Michalopoulos and Papaioannou (2014) pertains to the effects of institutions on the probability a pixel is not completely dark – a probability that has no clear economic meaning. By contrast,

[10]Examples where nightlights are used in highly aggregated form include Pinkovskiy and Sala-i-Martin (2016) and Martinez (2022). In addition the issues with nightlights that we emphasize here might not arise in research that focuses on cities or large settlements (e.g., Storeygard (2016) and Bluhm and Krause (2022)).

the results using the new consumption estimates are easy to understand. Using the rule of law variable, for example, the empirical models suggest that going from the lowest to the highest level of this variable leads to more than a 30 percent increase in consumption per capita using purchasing power parity dollars. The re-analysis of these two papers therefore illustrates how the new consumption data, adjusted to eliminate NCME, not only makes possible clear substantive interpretation of regression results, but also to avoid the very challenging problem of bias inherent when nightlights are used in spatially disaggregated econometric models.

The paper is organized as follows. Section 2 introduces the mathematical framework we employ to translate an asset index into a measure of consumption, describes the construction of this variable, and presents evidence to validate it. Section 3 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. Section 4 presents the consumption data for 42 countries over time and provides evidence that these data are capturing within-country variation in a meaningful way. Section 5 discusses non-classical measurement error (NCME) when using the spatially disaggregated data in regression. Section 5.1 focuses on NCME when using nightlights, and section 5.2 describes the problem of non-classical measurement error in the new estimates of consumption, as well as our approach for ridding the data of this error. Section 6 uses the new consumption data (with the NCME eliminated), to re-estimate models in Michalopoulos and Papaioannou (2013) and Michalopoulos and Papaioannou (2014). Section 7 concludes. Additional information is provided in the Appendix.

2. 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 captures a key aspect of development, consumption. 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.[11]

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

[11] 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.

the section by describing evidence to validate the measure, including a critical examination of the key assumptions within the framework.

2.1. A mathematical framework for translating an asset index into a measure of consumption. 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.[12]

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 (based on individual ownership of motor scooters, automobiles, electrical appliances, indoor plumbing, etc.). As a practical matter, in developing parts of the world y^A is typically observed at the individual level (e.g., using the DHS surveys employed here) whereas y^C is typically not observed at the individual level, and only some aspects of its country-level distribution (such as its mean or its dispersion) are available (e.g., using the WB-PIP data). Our approach will use the country-level measures of y^C to transform the individual-level measures of y^A into individual-level measures of consumption.

We adopt a framework similar to Chen and 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.[13]

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

$$y^C_{ict} = y^*_{ict} + \epsilon^C_{ict}, \tag{1}$$

$$y^A_{ict} = \alpha_{ct} + \beta_{ct} y^*_{ict} + \epsilon^A_{ict}, \quad \beta_{ct} > 0, \tag{2}$$

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}.

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

[13] 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.

Equation (1) 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 (2) expresses the assumed relationship between an asset index, y^A, and y^*. This relationship is linear and is a function of the country and time specific parameters, α_{ct} and β_{ct}.

Using equation (2), define a new proxy for y^* that is based on the asset index:

$$\widetilde{y^*_{ict}} = (y^A_{ict} - \alpha_{ct})/\beta_{ct}, \tag{3}$$

which can be written as

$$\widetilde{y^*_{ict}} = y^*_{ict} + \widetilde{\epsilon_{ict}}, \text{ where } \widetilde{\epsilon_{ict}} = \epsilon^A_{ict}/\beta_{ct}. \tag{4}$$

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} are known, $\widetilde{y^*_{ict}}$ can be observed at the micro level provided y^A_{ict} is also observable.

To identify the parameters α_{ct} and β_{ct}, note that combining equations (1) and (2) yields

$$y^C_{ict} = \frac{y^A_{ict} - \alpha_{ct}}{\beta_{ct}} + (\epsilon^C_{ict} - \epsilon^A_{ict}/\beta_{ct}). \tag{5}$$

The asset index, y^A_{ict}, is measured in different units than y^C_{ict} and y^*_{ict}. 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). That is, we renormalize the variables so that $E(y^A_{ict}) = 0$ and $Var(y^A_{ict}) = \sigma^2_{y^A_{ct}} = 1$ for all c and t. Given that $cov(y^A_{ct}, \epsilon^A_{ct}) = \sigma^2_{\epsilon^A_{ct}}$, equation (5) therefore implies that

$$E(y^C_{ct}) = \mu_{y^C_{ct}} = \frac{-\alpha_{ct}}{\beta_{ct}}, \quad \text{and} \tag{6}$$

$$\begin{aligned} Var(y^C_{ct}) = \sigma^2_{y^C_{ct}} &= 1/\beta^2_{ct} Var(y^A_{ct} - \epsilon^A_{ct}) + Var(\epsilon^C_{ct}) \\ &= 1/\beta^2_{ct}(1 - \sigma^2_{\epsilon^A_{ct}}) + \sigma^2_{\epsilon^C_{ct}} \qquad (7) \\ &= 1/\beta^2_{ct} + (\sigma^2_{\epsilon^C_{ct}} - \sigma^2_{\epsilon^A_{ct}}/\beta^2_{ct}), \qquad (8) \end{aligned}$$

where $\sigma^2_{\epsilon^i_{ct}}$ denotes the variance of ϵ^i_{ct} for $i = \{C, A\}$.[14]

Note that in equation (8), $\sigma^2_{\epsilon^C_{ct}}$ is the variance of the measurement error in y^C_{ict} (see eq. 1) and $\sigma^2_{\epsilon^A_{ct}}/\beta^2_{ct}$ is the variance in the measurement error in the asset variable when it has been transformed by the parameters α_{ct} and β_{ct} into a measure of consumption, $\widetilde{y^*_{ict}}$, (see eqs. 3 and 4). Thus, if the variances of these two measurement errors were very similar, then the variance of $y^C_{ct} \approx 1/\beta^2_{ct}$. For this reason, Assumption B is very useful.

Assumption B. $\sigma^2_{\epsilon^C_{ct}} - \sigma^2_{\epsilon^A_{ct}}/\beta^2_{ct} \approx 0$.

Since we are assuming that a transformed asset index can serve as a proxy for consumption, it does not seem unreasonable to assume that the variance of the measurement error in this transformed proxy is similar to the variance of the measurement error in consumption itself. We of course cannot verify this assumption directly, but section 2.2 discusses a testable implication of what we should observe empirically when Assumption B does or does not hold, and section 2.4.1 provides evidence supporting the validity of this assumption.

As noted, Assumption B implies that $Var(y^C_{ct}) = \sigma^2_{y^C_{ct}} = 1/\beta^2_{ct}$. And by eq. (6), $\mu_{y^C_{ct}} = \frac{-\alpha_{ct}}{\beta_{ct}}$. Therefore, under Assumptions A and B it is possible to obtain expressions for α_{ct} and β_{ct} using only *country level* moments of y^C_{ct}:

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

[14] Note that $Var(y^A_{ct} - \epsilon^A_{ct}) = Var(y^A_{ct}) + Var(\epsilon^A_{ct}) - 2cov(y^A_{ct}, \epsilon^A_{ct}) = 1 - \sigma^2_{\epsilon^A_{ct}}$.

Equation (9) implies that consistent estimates for α_{ct} and β_{ct} can be derived from consistent estimates of $\mu_{y^C_{ct}}$ and $\sigma_{y^C_{ct}}$. 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 (e.g., DHS) and another that provides the first and second order moments of the country-level consumption distribution (e.g., WB-PIP).

2.2. 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 linear relationship with y^*, y^C_{ict} and y^A_{ict} are also linearly related (eq. 5). 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 2.4 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 $E_{ct}(\widetilde{y^*_{ct}}) = \delta_0 + \delta_1 \mu^*_{ct}$.[15] 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.

[15] 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 $E_{ct}(\widetilde{y^*_{ct}}) = 0 - \frac{\alpha_{ct}}{\beta_{ct}} = \mu^C_{ct} = \delta_0 + \delta_1 \mu^*_{ct}$.

Violation of Assumption B. Suppose that Assumption A holds. Define the proxy $\bar{y}^*_{ict}$ as an asset index that has been transformed using equation (10):

$$\begin{aligned}\bar{y}^*_{ict} &= \sigma_{y^C_{ct}} y^A_{ict} + \mu_{y^C_{ct}} \qquad (11)\\ &= \sigma_{y^C_{ct}}(\alpha_{ct} + \beta_{ct} y^*_{ict} + \epsilon^A_{ict}) + \mu_{y^C_{ct}} \text{ (by eq.2)}\\ &= (\alpha_{ct}\sigma_{y^C_{ct}} + \mu^*_{ct}) + \sigma_{y^C_{ct}}\beta_{ct} y^*_{ict} + \epsilon^A_{ict}\sigma_{y^C_{ct}}. \qquad (12)\end{aligned}$$

Focusing on eq. (11), note that since $E_{ct}(y^A_{ct}) = 0$, it follows that

$$E_{ct}(\bar{y}^*_{ct}) = \mu_{y^C_{ct}} = \mu^*_{ct},$$

indicating that the country-level average of the new proxy equals the average of the true consumption indicator, y^*_{ct}. Thus violating Assumption B does not result in biased country means of the transformed asset variable.

If Assumption B holds, $\mu^*_{ct} = -\alpha_{ct}\sigma^C_{ct}$ and $\beta_{ct} = 1/\sigma_{y^C_{ct}}$ (see equation 9), which implies that the intercept in equation (12) vanishes and the slope is equal to 1. If Assumption B does not hold, the intercept in equation (12), $\alpha_{ct}\sigma_{y^C_{ct}} + \mu_{y^C_{ct}}$, is in general different from zero, and the slope, $\sigma^C_{ct}\beta_{ct}$, is different from 1.

Since $y^C_{ict} = y^*_{ct} + \epsilon^C_{ct}$, the relationships described in the previous paragraph also hold if y^C_{ict} is used instead of y^*_{ct} in equation (12). It therefore follows that one can test the validity of Assumption B by regressing the (sample counterpart of the) transformed asset proxy $\bar{y}^*_{ict}$ on y^C_{ict} and then examining whether the intercept in that equation is zero and the slope is one. Situations where the slope significantly deviates from 1 and the intercept significantly deviates from zero would indicate substantial departures from Assumption B, and vice versa. This test is conducted in section 2.4 using LSMS data on assets and consumption.

2.3. Constructing the training variable: Cluster-level measures of consumption. This section describes the steps involved in combining asset surveys and macro data on consumption to construct the consumption measure used as a training variable. A more detailed explanation can be found in Appendix B.

Previous studies have used an asset index directly for training purposes (Jean et al. 2016, Yeh et al. 2020), so a potential alternative strategy could be to follow this approach and then transform the predictions into estimates of consumption. This approach, however, has important drawbacks. First, the post-prediction transformation is infeasible when the necessary country-level data are unavailable. Second, predicting consumption per capita rather than an asset index provides an additional validation mechanism: when aggregated, our predicted consumption should align with national estimates in the WB-PIP dataset. This direct validation is impossible when using an asset index as the training variable because there is no comparable national benchmark for asset-based measures. And third, a critical requirement for prediction models of spatially disaggregated economic well-being is that the training variable be comparable across time and space. When an asset index is used as the training variable, this comparability is achieved by pooling asset variables from all surveys and then running a principal components analysis on the pooled data. However, this approach limits the number of asset variables that can be used in construction of the training variable, reducing its variability and accuracy and narrowing its applicability. Additionally, creating an asset index by estimating a principal components analysis on pooled asset data requires the strong assumption that the economic value of specific assets is consistent across countries and over time. The approach outlined in this section avoids these drawbacks by creating a training variable denominated in log dollars (using 2011 PPP). This method enables the use of different assets across surveys while maintaining comparability, thereby addressing the constraints associated with using an asset index directly for prediction.

To create the training variable, 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).[16] 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

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

(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.[17] 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 2.1, equation (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.[18] 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 34,482 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.

2.4. Validating the training variable. This section assesses the validity of the framework described in section 2.1 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 2.2. The most direct way to assess these implications is to utilize surveys that have measures of both consumption and assets. Since the DHS surveys we use as the main source of asset indices in this paper have no measure of consumption, we begin by using the Living Standards Measurement Study (LSMS).[19]

2.4.1. 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

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

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

[19] 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.

than one survey exists in a country, we select the survey closest to the years 2012-13. The result is a dataset composed of seven countries.[20] For each household, we compute a (log) consumption index and a (log) asset-based index, transforming the asset index as described in section 2.3. 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 $\sigma_{y_{ct}^C}$.

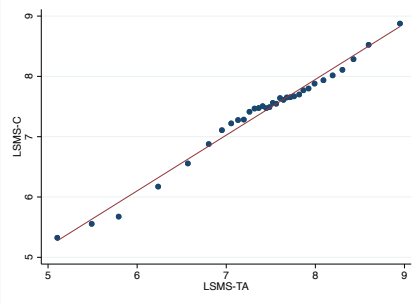
Given our focus on cluster-level data in the prediction exercises below, we focus on "cluster" level LSMS data, which is created using LSMS enumeration areas. In each enumeration area, we average the respondents' consumption scores and translated asset scores to create the cluster-level measures.[21] 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 are 2,455 clusters in the seven countries.

As underlined in section 2.2, Assumption A implies *there should be a* ***linear*** *relationship between* y_{ict}^C *and* $\widehat{y_{ict}^*}$. Panel (a) in Figure 2 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.

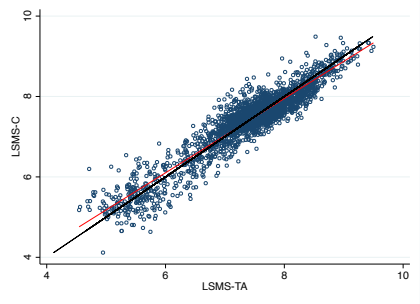
Assumption B implies that *if we regress the transformed asset proxy on* y_{ict}^C*, the intercept should be near zero and the coefficient for* y_{ict}^C *should be near one*. Panel (b) in Figure 2 shows the scatter plot of LSMS-C against LSMS-TA, along with the red regression line and the black 45-degree 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 fit is very good (the adjusted R-squared is 0.88), the constant is not significantly different from zero (coefficient is .56 and p=.18), and the slope is not statistically different from 1 (the coefficient is 0.92 and the p-value for the test that the slope equals to 1 is 0.15). In addition, the joint hypothesis of slope=1 and intercept=0 cannot be rejected at conventional significance levels. These results therefore suggest that Assumption B is not violated in this data.

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

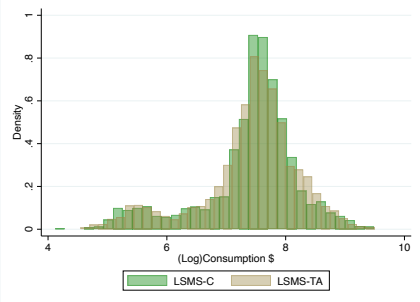
[21] Some LSMS enumeration areas have very few respondents, and so we impose a minimum of 10 households in a cluster for inclusion of the cluster. We have also eliminated 6 data points which are extreme outliers, all belonging to Ghana.



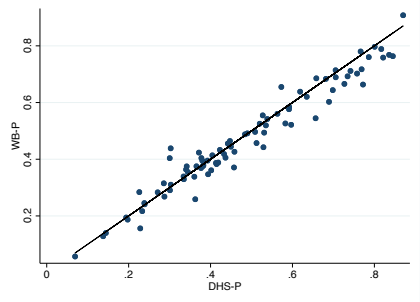
(a) Binned scatter plot of LSMS-C vs. LSMS-TA (both measured as log consumption dollars per capita)



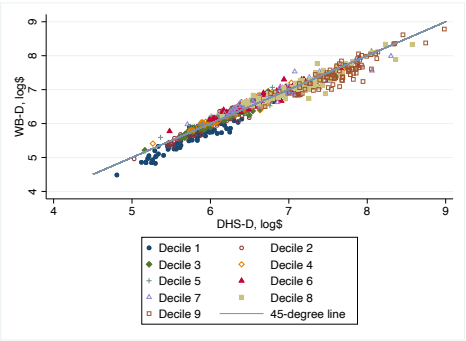
(b) Scatter plot of LSMS-C vs. LSMS-TA (both measured as log consumption dollars per capita), along with 45-degree line (in black) and regression line (in red)



(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-P) and DHS transformed assets (DHS-P) , 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 2. Validating the transformation of asset indices into measures of consumption.

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.

2.4.2. 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 2 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 2 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.

3. Prediction

This section lists the predictors (features) employed to estimate the prediction models (section 3.1); describes the algorithm and models employed (section 3.2); and assesses prediction accuracy (section 3.3). In addition, we aggregate cluster data to the national level and compare the resulting measures of national consumption with data from the World Bank (section 3.4).

3.1. Features. We use a total of 29 predictor variables, each defined over the 10 km^2 cells. Appendix C provides a detailed description and data sources.

(1) **Nightlights.** Satellite-based measures of cell-level intensity of light at night. There are two main sources of satellite data. 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. Since our data spans the transition from DMSP to VIIRS, for prediction, we use the harmonized version of nightlights (which aims to make the DMSP- and VIIRS-based satellite data comparable) described in Li et al (2020).

(2) **Other time varying features.** Population density, CO2 production and malaria incidence.

(3) **Features related to geography.** Ecosystem type, ruggedness of terrain, elevation, and caloric yield of land.

(4) **Features related to cell location.** Distance from the cell to the capital, a highway, the coast, a harbor, a river, any Christian mission, a catholic mission, and a protestant mission. Also the cell's latitude and longitude.

(5) **Features related to climate**: Cell average temperature and rainfall, as well as year-specific deviation from this average.

3.2. 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.[22]

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.[23]

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.

3.3. 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.

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

[23] We use Ahrens and Schaffer's (2022) "pystacked" package in Stata, which implements the random forest model using scikit-learn's "sklearn.ensemble.StackingRegressor." Further information is found at https://statalasso.github.io/.

For each omitted survey s, we produce out-of-sample forecasts and compute the MSE and the R^2. 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.[24]

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 three models. The first row, model RF-1, reports the result for a random forest model that includes only NIGHTLIGHTS as a predictor. The median MSE is 0.200 and the median R^2 is 0.611. The second line of the table presents our full model, RF-2, where the predictors include NIGHTLIGHTS along with the 28 other variables discussed above. Prediction accuracy improves substantially: the median MSE, for example, is 47 percent larger in RF-1 than in RF-2. Thus, the additional predictors are clearly improving accuracy.[25]

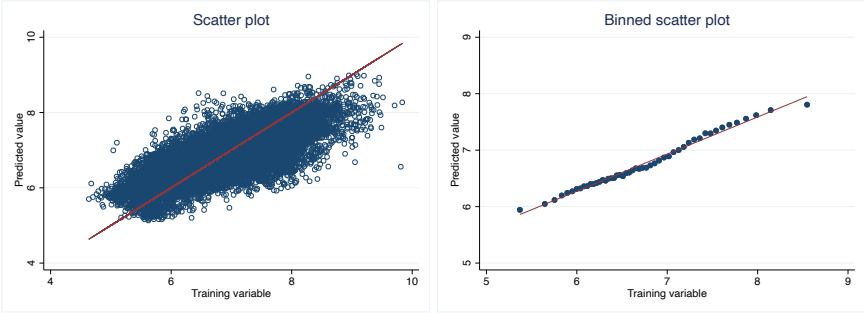
Predictors included	Median MSE	Median R^2
RF-1: NL ONLY	0.200	0.611
RF-2: NL PLUS ALL OTHER PREDICTORS	0.136	0.699
RF-3: ALL PREDICTORS EXCEPT NL	0.141	0.680

Table 1. PREDICTION ACCURACY. This table displays the median MSE and R^2 from the 85 out-of-sample predictions for each model. The models are estimated using a random forest algorithm and differ in the predictors included in the model.

Panel (a) in Figure 3 displays a scatter plot showing the relationship between the training variable and the (out-of-sample) predicted values from RF-2. The plot demonstrates that the overall correlation is quite strong, indicating that the model captures the general pattern very well. Panel (b) provides a binned scatterplot that smooths out individual variability, offering a clearer view of the underlying relationship between these variables. This plot reveals that the relationship is highly linear across most of the distribution, with small deviations appearing

[24] We prefer the MSE to the R^2 as a goodness of fit measure because the (out of sample) R^2 focuses exclusively on the squared correlation between the training and the predicted variables, and therefore is unaffected by systematic biases that the predictions might contain (i.e., the predictions y_1 and $y_1 + c$ will have the same R^2 regardless of the value of c). The MSE, a function of the average distance between the training variable and the prediction, captures any such biases. In our data the correlation between the MSE and the R^2 in our core model is -0.42.

[25] We also estimated a KNN model with only nightlights as a predictor and this model produced slightly better predictions than RF-1. The median MSE was 0.18 and the median R^2 was 0.627.



(a) Training data vs predicted value, scatter plot

(b) Training data vs predicted value, binned scatter plot

Figure 3. PREDICTIONS VERSUS TRAINING VARIABLE. Panel (a) plots a scatter plot of the out-of-sample predictions (RF-2) against the training variable; Panel (b) plots the bineed scatter plot relating these variables.

at the two ends. Specifically, the predictions tend to underestimate the values of the very rich and overestimate the incomes of the very poor.

Which predictors are contributing the most to prediction accuracy? The first column in Table 2 shows the ten most informative predictors (based on mean decrease in impurity) using RF-2. The most important variable by a substantial margin is NIGHTLIGHTS, followed by CO2 and POPULATION DENSITY. Variables tapping a cell's proximity to the capital, to highways and to Christian missions are also important. It is interesting to note that four of the top ten variables (the first three and MALARIA) are predictors that vary over time. The second column shows variable importance for RF-3, the model that excludes NIGHTLIGHTS. The set of variables is essentially the same as those in the first column. Thus, we see that NIGHTLIGHTS is an important predictor that should be included in any model, but that other variables contribute to prediction accuracy.

The prediction accuracy of RF-2 described in Table 1 is similar to that of Yeh et al. (2020), though the results are not directly comparable given differences in the training variable and in the set of DHS surveys used. Appendix D.2 describes a direct comparison of our approach with that of Yeh et al. (2020). The analysis shows that the RF approach with the rich set of predictors can yield more accurate estimates than the Yeh et al approach. More specifically, using Yeh et al's training variable and sample, the error from Yeh et al's KNN model is 20% higher than the error using RF-2, and the error using Yeh et al's CNN algorithm is 13% higher than the error from the RF-2 model. This improvement using RF-2 is obtained at a much

	Relative Variable Importance	
Ranking	RF-2	RF-3
1	Nightlights (.24)	CO_2 (.22)
2	CO_2 (.13)	Population Density (.14)
3	Population Density (.09)	Distance to capital (.07)
4	Distance to capital (.05)	Distance to Christian mission (.05)
5	Distance to Christian mission (.04)	Distance to highway (.05)
6	Distance to highway (.04)	Latitude (.04)
7	Malaria (.04)	Malaria incidence (.04)
8	Latitude (.03)	Longitude (.03)
9	Average Rain (.03)	Average Rain (.03)
10	Distance to protestant mission (.03)	Distance to protestant mission (.03)

Table 2. VARIABLE IMPORTANCE. This table provides the 10 most important predictors for models RF-2 and RF-3. Relative importance based on mean decrease in impurity are given in parentheses.

lower computational cost. The CNN algorithm in Yeh et al. (2020) combined with their use of daylight imagery is very costly computationally, substantially limiting the scalability of their approach. By contrast, the approach here generates the over 4 million cell estimates discussed below in under 13 minutes using a laptop computer.

Why is the low computational cost important? It is important to bear in mind that our approach does not require the use of a specific set of predictors. That is, we are not arguing that the 29 predictors included in RF-2 are the best and only way to predict our training variable. Instead, we are arguing that an RF model with a rich set of relevant predictors that includes NIGHTLIGHTS should provide accurate predictions. Thus, our approach could be implemented with different predictors in different contexts. This is valuable because it allows researchers to easily adapt our approach outside of sub-Saharan Africa, or inside sub-Saharan Africa after the time period here.

To what extent, then, does prediction accuracy vary with the set of predictors? To address this question, we have re-estimated RF-2 eleven times, each time omitting one of the most important variables that appears in Table 2. We have calculated the MSE not only of the median survey, but also of the survey at the 25th and at the 75th percentile. The results are presented in Figure 4.

The figure shows that no matter which one of the most important predictors is excluded – including NIGHTLIGHTS – the prediction accuracy scarcely varies at all. For example, the 25th

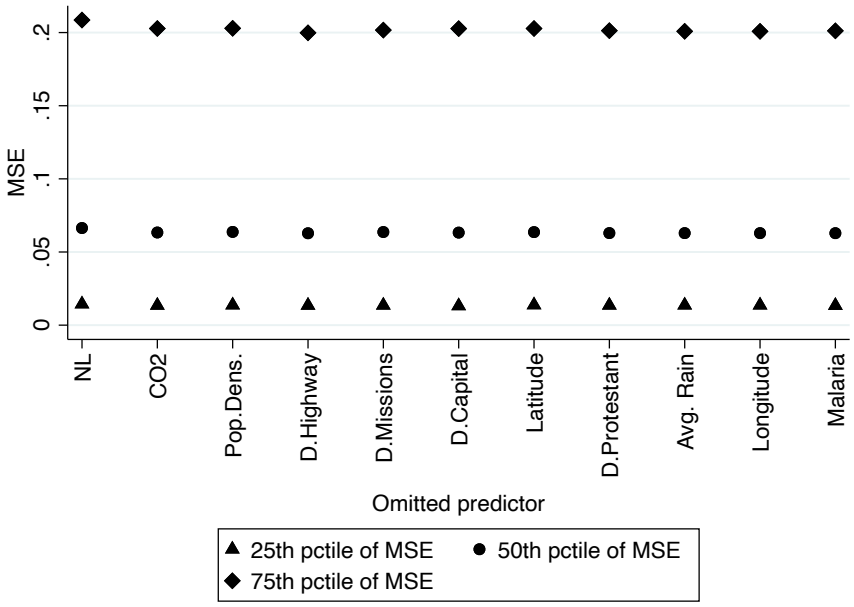


Figure 4. Out-of-sample prediction accuracy when one of the most important predictors is excluded. This figure provides the value of the MSE for the locations at the 25th, 50th and 75th percentiles in the distribution of MSE. The values are based on 85 out-of-sample sets of predictions when RF-2 is estimated without the variable listed on the x-axis.

percentile of the MSE values ranges from 0.0998 (when average rain is omitted) to 0.1091 (when nightlights is omitted). And the 75th percentile of the MSE values ranges from 0.1723 (when distance to the capital is omitted) to 0.1846 (when nightlights is omitted). In addition, the mean MSE across all 85 locations varies minimally: this mean is highest (0.1692) when nightlight is omitted, but this mean is only 0.0056 higher than the lowest mean MSE (which occurs when distance to a highway is omitted). Thus, though excluding nightlights is ill advised, prediction accuracy is highly robust to permutations in the set of predictors used in the model. In section 6 we provide a specific example of why this is valuable.

3.4. Predictive performance of national consumption. Since our training variable measures consumption, there is a second pathway for evaluating predictive performance, which is to use the DHS cluster weights to aggregate the cluster-level predictions for each survey to a national-level measure of log consumption. We can then compare these national-level measures with the corresponding WB-PIP measures of national log consumption. If prediction accuracy is high, our method can provide a simple and cheap way of obtaining consumption per capita

at different levels of aggregation, including for country-years for which WB-PIP data do not exist.

Figure 5 compares the country-level per-capita consumption estimates from WB-PIP with those from RF-2, alongside the 45-degree line for reference. The correlation between these two sets of estimates is strong (0.85), and most points lie close to the line, indicating that RF-2 predictions generally align well with WB-PIP estimates. However, this alignment is tighter for countries in the middle of the income distribution than for those at the extremes. This pattern reflects the individual-level tendencies discussed earlier, where the model overestimates consumption at the very low end of the distribution and underestimates it at the high end. Consequently, at the country level, RF-2 tends to overpredict consumption in very poor countries, such as the Democratic Republic of the Congo (the poorest country in our sample), and underpredict consumption in the richest countries, exemplified by Namibia and Ghana, the two outliers in the upper-right corner of the figure.

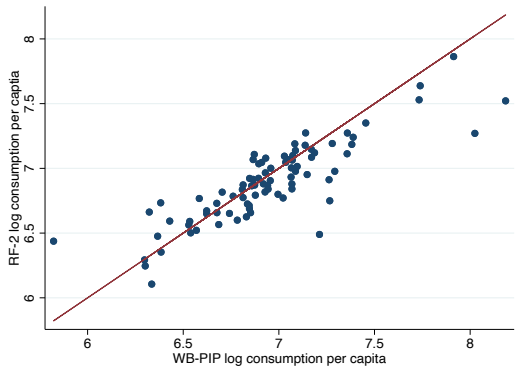


Figure 5. LOG CONSUMPTION P.C AT THE COUNTRY-YEAR LEVEL: RF-2 (Y-AXIS) VS. WB-PIP (X-AXIS). This figure displays country-level (log) consumption per capita from WB-PIP and using the out-of-sample predictions from RF-2 for the 85 country-years in our sample. The red line is the 45-degree line.

4. Consumption data for sub-Saharan Africa

We now proceed to the last step in data creation, which is to develop consumption estimates for all grid cells in sub-Saharan Africa over time. We do this by using all the training data 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. Figure 6 shows the population-weighted distribution of the cell estimates. The histograms include all grid cells (over 4.1 million)

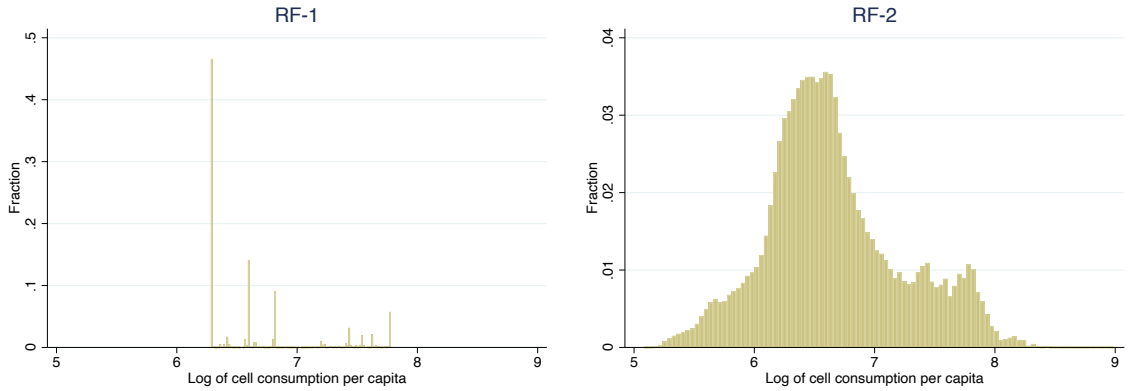


Figure 6. THE POPULATION-WEIGHTED DISTRIBUTION OF CONSUMPTION. The left panel depicts the population weighted distribution of consumption from RF-1 (where NIGHTLIGHTS is the only predictor. The right panel presents this distribution for the unadjusted estimates of log consumption using RF-2, the prediction model with the full set of predictors, including NIGHTLIGHTS. The data include over 4.1 million cells from 43 countries for the period 2003-18.

across 43 countries from 2003-2018. For comparison purposes, the left histogram presents the consumption distribution resulting from RF-1, the model using only NIGHTLIGHTS. The distribution from this model is concentrated on a very small range of values, with a huge proportion of values in the same narrow band. This of course is due in large part to the absence of nightlight in many cells. Using the consumption estimates (the right panel), there is much more variability in the estimates and much higher maximum values. The histograms therefore clearly illustrate that prediction models using only nightlights 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 consumption estimates that exhibit the rich variability one 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 consumption estimates is aligned with variation in estimates from an external dataset that exists at the subnational region level. The 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).[26] We are particularly interested in comparing the 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 our estimates of regional consumption within countries.

The HDI and consumption data provide estimates with unknown biases and unknown measurement error, so this comparison should not be viewed as a validation exercise of the consumption data. But since each dataset is computed using different data sources and different methodologies, if the consumption and HDI 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 of consumption are capturing within-country variation across cells in a meaningful way.

To make the comparison, we aggregate the consumption measures in each cell to the level of the HDI regions.[27] 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.

	HDI Income (1)	HDI Education (2)	HDI Life Exp. (3)
Consumption, RF-2	0.81	0.65	0.39

Table 3. WITHIN-COUNTRY CORRELATION USING DATA AT THE LEVEL OF SUBNATIONAL REGIONS. The cells display within-country correlations – which are based on deviations of the variables from their country means – between RF-2 consumption estimates and the subnational HDI variable named at the top of each column. There are 244 regions from 20 countries.

Table 3 provides the "within-country" correlations of our consumption measure with the HDI variables. These within-correlations are calculated using the deviations of each region from its country mean. We find that the estimate of regional consumption based on RF-2 has a correlation of 0.81 with the HDI measure of income per capita. This is stronger than the within-correlation for the other two components of the HDI. Thus, Table 3 indicates that the

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

[27] 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.

regional variables created by aggregating consumption cells to the regional level are strongly correlated within countries to the HDI data that measure income but less strongly correlated with the other variables tapping regional well-being.

5. The problems when using nightlights and predicted consumption in regression analysis

We have shown that estimates of consumption from RF-2 provides an accurate measure of consumption per capita. It can therefore be harnessed to pursue a variety of distinct objectives. One is to provide detailed, ground-level descriptions of economic well-being, 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 consumption estimates as proxies for economic well-being in regressions using a spatially disaggregated unit of analysis. In regression settings, stricter standards are necessary to address potential concerns about measurement error. This will be our focus in what follows.

This section begins by developing the argument outlined in the Introduction, which serves as a key motivation for this paper: nightlights data, when used at a highly disaggregated level, are particularly susceptible to large measurement error. More importantly, this error is non-classical, that is, it is correlated with economic well-being. As is well-known, non-classical measurement error is particularly problematic when standard estimation techniques are employed. First, coefficients can be biased regardless of whether the variable measured with error is used as independent or dependent variable. Second, results are difficult to interpret, as biases can be either attenuating or amplifying. Finally, avoiding these problems with standard techniques such as using instrumental variables is quite challenging (see below). As we will illustrate in section 6, the biases can be so severe that correcting for them may completely reverse the results.

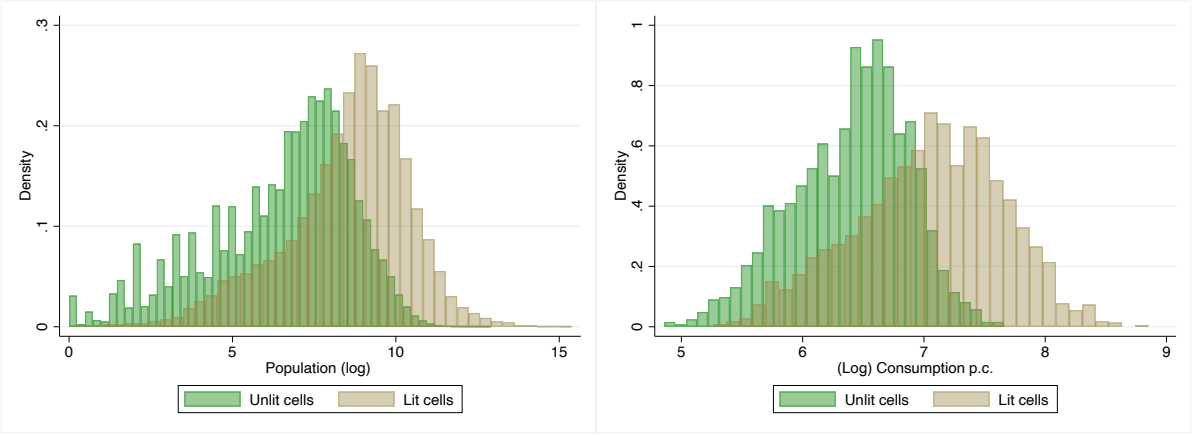
If we address this limitation with nightlights by using the consumption estimates, we face a similar issue because these estimates also suffer from non-classical measurement error. In the case of the consumption estimates, however, it is possible to eliminate the non-classical measurement error. Below we first describe the problem of non-classical measurement error

in nightlights. We then turn to the consumption, introducing a simple procedure designed to eliminate the non-classical component, leaving only classical measurement error, a much more manageable type of error in regression.

5.1. Nonclassical measurement error in NL. 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, 2021) for a comprehensive review). Here we focus on a problem that stems from the fact that in large areas of the developing world and, in particular, in most of Africa, 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 88% of the 10x10km cells are dark in 2018. We argue in this section 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.

Why should we expect that the "problem of darkness" leads to non-classical measurement error in NL? To answer 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.

A central reason that NCME arises with nightlights is because the vast majority of cells in Africa are dark. Importantly, the dark areas are not void of people or economic activity. Panel (a) of Figure 7 shows the distribution of estimated cell population (using Gridded Population of the World data) across lit and dark 10km^2 cells in 42 African countries in 2018, when the more accurate VIIRS satellite data is available. Although lit cells are obviously more populated, there is a large overlap in the two distributions. In 2018, the available population estimates suggest that 51.4% percent of the population lived in dark cells.



(a) Population in lit and unlit areas, Consumption 2018

(b) (Log) Consumption p.c. in lit and unlit areas, DHS training data

Figure 7. Relative frequency of (log) population (**Panel a**) and (log) consumption per capita (**Panel b**) in lit and unlit areas in Sub-Saharan Africa. The histogram in panel (a) reflects the relative frequency distribution of population in lit versus unlit areas using all cells in the 42 countries of the consumption data for the year 2018. Panel (b) presents the histogram in lit versus unlit areas of the training variable in all DHS clusters in the training data set (see section 2 for details). Nightlights data comes from the VIIRS satellite.

One could argue that, because of the way population is attributed to cells in the Gridded Population dataset, the population in dark areas is overestimated.[28] However, if we focus instead on economic activity a similar picture arises. Panel (b) of the same figure shows the distribution of (log) consumption per capita in dark and lit cells using only our training data, and thus using only surveyed areas. Although 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.

The link between the "problem of darkness" and NCME in nightlights 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

[28]The Gridded Population of the World used in Panel (a) of Figure 7 computes gridded population data by redistributing population counts from census and administrative units to a global raster grid using a uniform areal weighting method, assuming an even population distribution within each unit. See https://sedac.ciesin.columbia.edu/downloads/docs/gpw-v4/gpw-v4-documentation-rev11.pdf for details.

$y^* = 1$ 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 must obviously be 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 almost 90% of the territory and, as is clear from Figure 7, there are many people and much economic activity in these dark spaces. This implies that $u = -y^*$ 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 we will present in section 6, 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:

$$\delta = \hat{\beta} - \beta = \frac{cov(x'u)}{var(x)} + o_p(1). \tag{13}$$

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 $\delta \overset{p}{\nrightarrow} 0$. 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

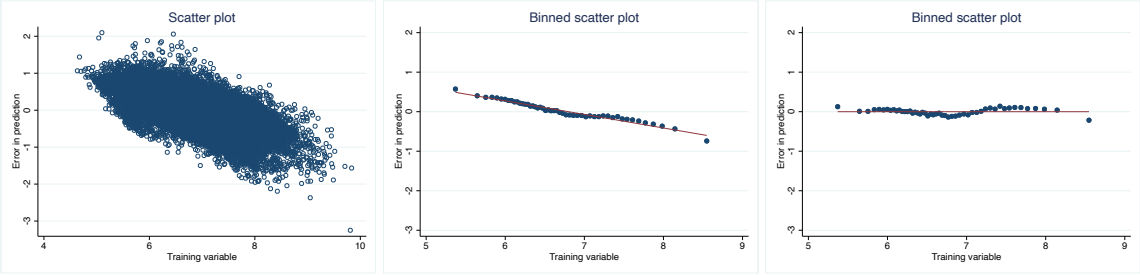
$$\hat{\beta} = \Big(1 - \frac{cov(y^*, u) + \sigma_u^2}{\sigma_{NL}^2}\Big)\beta + o_p(1) = (1 - \gamma)\beta + o_p(1). \tag{14}$$

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.[29] 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 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.[30]

[29] 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.

[30] See Schennach (2016) for a review of the nonclassical measurement error problem and solutions.



(a) Prediction error vs. training data

(b) Prediction error vs. training data

(c) Prediction error from transformed data, vs. training data

Figure 8. 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.

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.

5.2. Nonclassical measurement error in consumption estimates: A simple solution. For the reasons outlined in Section 5.1, any alternative data source intended to replace NL in regression analysis must not only be accurate, but also free of non-classical measurement error. Unfortunately, we cannot assume that the consumption estimates meet this condition. Panel (a) of Figure 8 presents the predicted errors from RF-2 plotted against the training data, revealing a negative correlation between prediction errors and the values of the training variable: as seen in Figure 3, the model tends to overestimate incomes at the lower end and underestimate those at the upper end. Panel (b) displays a binned scatter plot confirming a linear, negative relationship between these variables. As a result, the consumption estimates themselves suffer from non-classical measurement error, posing the same types of problems previously discussed in the context of using nightlights data in regression.

Although a detailed examination of how to incorporate machine learning-generated regressors into regression frameworks is beyond the scope of this paper, it is still possible to propose a straightforward solution that addresses the nonclassical component of the

measurement error.[31] In particular, we can leverage the training sample as an auxiliary data source to derive a simple transformation of the consumption estimates, ensuring they are free of nonclassical error. The transformation is constructed as follows. Consider a target variable y for which a proxy $\tilde{y}$ exists that might contain non-classical measurement error, v. 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,[32]

$$\tilde{y} = \alpha_0 + \alpha_1 y + \epsilon. \tag{15}$$

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

$$\tilde{y}^T = \frac{\tilde{y} - \alpha_0}{\alpha_1} = y + \epsilon/\alpha_1. \tag{16}$$

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 3 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 8 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 4 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 4 display the mean, standard deviation, 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

[31] This remains an active area of research, and the available results are still incipient. For recent references, see Battaglia et al. (2024), which addresses the issue specifically in settings where generated variables serve as regressors. In our illustration, however, our consumption estimates will be used as the dependent variable.

[32] Panel (c) in Figure 8 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.

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 almost 50% larger).

The adjustment described here can always be applied to proxies generated via supervised machine learning when it is reasonable to assume that the training variable is a representative sample of the target variable. But the table illustrates a clear tradeoff between the adjusted and unadjusted estimates. The unadjusted estimate is more accurate (and therefore more useful for descriptive purposes) but suffers non-classical measurement error. The adjusted estimate has more measurement error, but the error is classical rather than non-classical. The adjusted estimate is therefore more suitable in regression analysis.

	Mean	Std	min	max	$corr_{(\hat{e},\widehat{y^*_{rct}})}$	$corr_{(\hat{y},\widehat{y^*_{rct}})}$	MSE
	[1]	[2]	[3]	4	5	6	7
$\widehat{y^*_{rct}}$	6.74	0.699	4.634	9.835			
$\hat{y}_{RF2}$	6.74	0.564	5.132	9.012	-0.592	0.815	0.136
$\hat{y}^T_{RF2}$	6.74	0.858	4.262	10.168	0	0.815	0.192

Table 4. 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 $\hat{y}^T_{RF2} = (\hat{y}_{RF2} - 2.33)/.657$, where these coefficients have been computed in a regression of $\hat{y}_{RF2}$ on $\widehat{y^*_{rct}}$.

6. Illustration: Institutions and economic development

This section uses the (transformed) measure of consumption to revisit 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 a spatially disaggregated measure of nightlights as the dependent variable to study the effects of institutions on development. We re-estimate the empirical models from these two papers, substituting the adjusted measure of consumption, $\hat{y}^T_{RF2}$, for nightlights. The resulting conclusions about the role of institutions in development are fundamentally different than those found in the original MP papers.

We use this analysis to make two points about the use of our consumption measure in the place of nightlights. First, since the estimates we create are measured in consumption

dollars, they allows researcher not simply to identify the presence of significant relationships but also to quantify in substantively meaningful terms – dollars per capita – 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. Second, we argue that non-classical measurement error that is present in NL – but not in $\hat{y}^T_{RF2}$ – explains why the substantive results change so much when we change the outcome measure from nightlights to consumption, leading to coefficient estimates on institutions that have a positive bias in one MP paper and negative bias in the other. This discussion underlines why the non-classical measurement error in NL can often pose difficult challenges with interpreting regression results whenever nightlights are used as a measure of development in spatially disaggregated regressions.

6.1. Re-estimating models of institutions and development using estimated consumption.

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).[33] 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 – typically implemented as indicator variables for whether a pixel is lit – are used as a proxy for economic development. To replicate the MP13 and MP14 analyses with the consumption 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 consumption data includes sub-Saharan African countries, which means we do not have data for five countries in the MP data.[34] 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 E in the Appendix presents results from additional empirical models.

6.1.1. 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

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

Panel B of their Table IV (p. 177) and, for convenience, they are reproduced in Panel A of Table 5. These models exploit the fact that ethnic group boundaries often transcend national 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 5 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 5 re-estimates the same models as Panel A using the subset of pixels that are common to the consumption 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 fixed-effects specifications considered in MP14.

Panel C re-estimates the same models using the measure of log consumption from the full 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.[35] Section E in the

[35]The difference between the results using nightlights and results using the consumption 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 very 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)
Adjusted 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	0.4425**	0.1827***	0.3204***	0.1126**				
	(0.1847)	(0.0700)	(0.1181)	(0.0556)				
					0.6035***	0.2479***	0.4487***	0.1923***
					(0.1655)	(0.0860)	(0.1104)	(0.0699)
Adj. R-squared	0.193	0.789	0.416	0.815	0.278	0.790	0.459	0.818
N	40872	40872	39251	39251	40872	40872	39251	39251
	Panel D: Dep. variable is log consumption p.c., RF without capital distance (Reduced Sample)							
RULE OF LAW	0.4553**	0.1505***	0.3466***	0.1208**				
	(0.1879)	(0.0536)	(0.1210)	(0.0512)				
CONTROL OF CORRUPTION					0.6266***	0.2216***	0.4814***	0.1972***
					(0.1681)	(0.0661)	(0.1141)	(0.0607)
Adj. R-squared	0.193	0.819	0.400	0.832	0.283	0.821	0.446	0.835
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 5. 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-D 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.

Appendix provides results from estimating other models in MP14 using nightlights and the consumption 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, with the results for RULE OF LAW being particularly robust.[36]

[36]See Appenix E for details.

Robustness to a different prediction model. In section 3.3 we demonstrated that the prediction accuracy of the random forest models is largely unaffected by the removal of any specific predictor, especially when nightlights are included (see Figure 4 above). Thus, the consumption estimates should also be very similar when using variants computed with different subsets of features, as long as the feature set is sufficiently comprehensive. This implies that the results of regression models using estimated consumption should also be robust to small permutations in the set or random forest predictors used to generate the estimates, although the precision of coefficient estimates might decrease if larger errors are introduced into the prediction model.

This robustness is useful in the re-analysis of the MP14 data: one of the key findings in MP14 is that the effect of national institutions on development is heterogeneous across geographic areas. The analysis in MP14 Table VIII, for example, shows that pixels near the capital may benefit from strong national institutions, but this effect disappears in pixels that are from the capital. Given that distance to the capital is one of the variables used to compute consumption, a potential concern might be that the positive correlation observed between consumption and national institutions could arise from their shared correlation with distance to the capital, rather than from the direct relationship between national institutions and development.

A straightforward way to address this concern is to re-train the prediction model without including distance to the capital as a predictor. As we discussed with respect to Figure 4, excluding a predictor has a negligible effect on the median MSE in the DHS training data. And the correlation between the predictions from the two models (including and excluding distance to the capital) is 0.996, and the correlation between the errors of the two models is 0.989.[37]

How does removal of capital distance from the RF prediction model affect the results of the MP14 models? We re-created the consumption estimates using a random forest model without capital distance and then transformed these estimates to eliminate non-classical measurement

[37] Additionally, the correlation between distance to the capital and the predictions from the full RF-2 model (r= -0.314) is very similar to the correlation of this distance and the prediction from the RF model excluding distance from the capital (-0.283).

error following the procedure outlined in section 5.2. Panel D of Table 5 presents the results when these new estimates of log consumption are used in the MP14 models. Focusing on models 4 and 8 – the models with fixed effects as well as the pixel-level controls like capital distance – the results for the coefficients on the institutional variables are slightly larger than those in Panel C, but are nonetheless quite similar, and the substantive conclusions we draw about the relationship between institutional quality development is not affected by which prediction model we employ to produce the consumption variable. Thus, the results in Panel C cannot be attributed to the inclusion of capital distance in the prediction model.

Table E.4 in Appendix E explores whether the heterogenous effect identified by MP14 still holds when the consumption estimates are employed in place of NL. Consistent with the MP14 nightlights models, the estimated effect of institutional variables are larger for the set of pixels that are close to the capital – roughly twice as large – than for the set of pixels that are far from the capital. But unlike the results from the MP14 nightlights models, the coefficients for the pixels that are far from the capital are positive and reasonably precisely estimated. Identical conclusions are reached regardless of whether distance to the capital is included in the prediction model for computing consumption.

In sum, the evidence here indicates that the weak relationship in MP14 between national institutions and development is quite likely due to the use of nightlights as a development proxy. The conclusions about institutions shift dramatically when the consumption variable is used as the development proxy, revealing that institutions, especially the rule of law, have a large positive impact on development.

6.1.2. 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 6, Panel A. The dependent variable in the first five columns of their table, our focus here, is a dichotomous measure of nightlights,[38] and

[38]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 E in the Appendix).

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.[39] The main result is that JURISDICTIONAL HIERARCHY is positive and significant at the 5% level.

	(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	-0.0219	-0.0167	-0.0257	-0.0209	-0.0209
	(0.0661)	(0.0267)	(0.0278)	(0.0225)	(0.0215)
Adj. R-squared	0.001	0.761	0.774	0.814	0.820
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 6. 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 consumption estimates 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 6 re-estimates the models using the subset of pixels that are common to the consumption data 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 $\hat{y}^{T}_{RF2}$ 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 E 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 measure of consumption.

[39] 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.

6.1.3. Substantive interpretation of the results. A central motivation for developing the spatially fine-grained estimates of log consumption is that they permit a clear substantive interpretation of regression results that nightlights do not. To illustrate why this is valuable, it is useful to interpret the coefficients from the MP regression tables presented above. Consider the results from Table 5. The dependent variable in the nightlights models is an indicator variable describing whether a variable is lit or completely dark. Using the estimates from column (4) in panel B, going from the worst to the best value of RULE OF LAW implies that the probability that a pixel is lit increases by 10.4 percentage points. It is challenging to interpret this increase because it is impossible to put a specific value on the economic benefit of residing in a lit pixel. Using the measure of log consumption eliminates this problem because the outcome variable is denominated in log dollars. Using the estimates from column (4) in Panel C, going from the worst to the best value of RULE OF LAW leads to 31.7 percent increase in consumption per capita, and to an increase of 34.0 percent using the Panel D results from model 4. To put it somewhat differently, if Democratic Republic of Congo (which has a RULE OF LAW score of -1.88) could achieve Botswana's level of RULE OF LAW (which is 0.615), then the probability a DRC pixel is lit would increase by 9.2 percentage points and the consumption per capita would increase by 28.1% (using panel C). In 2018, WP-PIP reports that the mean consumption per capita was 626 dollars in the DRC. The results therefore suggest that if the DRC had Botswana's level of RULE OF LAW, its mean consumption would increase by 176 dollars.

For MP14, we find no robust effect of JURISDICTIONAL HIERARCHY on consumption. But if we use the nightlights results from model 5 in Panel B, going from the lowest to highest score of this institutional variable would increase the probability that a pixel is lit by 6.9 percentage points. As noted in the previous paragraph, it is very difficult to interpret this effect substantively. Using the results from model 5 in Panel C, going from the lowest to highest score of JURISDICTIONAL HIERARCHY would *decrease* consumption by 8.4 percent. Thus, the magnitude of the effect of JURISDICTIONAL HIERARCHY on consumption is a small fraction of the effect of RULE OF LAW or CONTROL OF CORRUPTION on consumption.

These results in Tables 5 and 6 therefore illustrate nicely the value *in terms of interpretability* of the spatially disaggregated measures of consumption. Compared with nightlights, where

the economic value of a lumen is unknown and almost certainly varies across time and space, the consumption metric is familiar, substantively meaningful, and comparable across time and space.

6.2. Explaining the divergence of the results: Non-classical measurement error in NL. The divergent results across panels in Table 5 and in Table 6 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 consumption 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 non-classical measurement error in nightlights?

As described in section 5, 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. Recall that u and y^* (the "true" measure of development) are negatively related as discussed in section 5. Therefore, if the national institution variables are in fact correlated with development, as Panel C in Table 5 strongly suggests, 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.

More concretely, in the MP14 data 89.5% of pixels are unlit in the data used in model 4 of Table 5. Thus, the relationship between the NI variables (rule of law and control of corruption) and nightlights will depend on how these variables are related to the probability a pixel is lit. Although this probability is essentially zero, Table 7 shows that when we estimate the core fixed effects models (4 and 8 in Table 5) using only dark or only lit pixels, the results show a strong relationship between all the national institutions variables and consumption. In other words, the extensive classification error in the nightlights data obscures the relationship between NI and NLs, preventing the data from capturing the positive, comparable association that exists in areas that remain dark.

	(1)	(2)	(3)	(4)
	Dark cells		**Lit. cells**	
RULE OF LAW	0.0934*		0.2080***	
	(0.0485)		(0.0474)	
CONTROL OF CORRUPTION		0.1605***		0.2662***
		(0.0570)		(0.0478)
Adj. R-squared	0.865	0.867	0.712	0.715
N	35109	35109	4142	4142
Ethnicity fixed effects	Yes	Yes	Yes	Yes
Population density and area	Yes	Yes	Yes	Yes
Location controls.	Yes	Yes	Yes	Yes
Geographic controls	Yes	Yes	Yes	Yes

Table 7. NATIONAL INSTITUTIONS AND ECONOMIC DEVELOPMENT IN LIT AND DARK CELLS. Models 1 and 2 re-estimate models 4 and 8 from Panel D in Table 5 using only dark cells. Models 3 and 4 re-estimate models 4 and 8 from Panel D in Table 5 using only lit cells. * indicates p<.10, **indicates p<.05, and *** indicates p<.01.

Consider next the MP13 results on the relationship between JURISDICTIONAL HIERARCHY (JH) and development. The discussion in section 5 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 areas with high concentrations. 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.[40]

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.[41] As a result, this amplification bias can lead to findings of a positive effect 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.

7. 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

[40] 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.

[41] 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 6), but it is unlikely to solve the bias problem described in this paragraph.

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, thereby leading to unwarranted conclusions about the causes of development. 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 meaningful way. Thus, the substantively interpretable metric of the consumption estimates, 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.

8. References

Acemoglu, D., T. Reed, and J. A. Robinson (2014) ": Economic development and elite control of civil society in Sierra Leone," *Journal of Political Economy* 122, 319–368.

Ahrens, A., Hansen, C.B. and Schaffer, M.E., "Pystacked: Stacking generalization and machine learning in Stata." arxiv link, url: https://arxiv.org/abs/2208.10896

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.

Battaglia, L., T. Christensen, S. Hansen and S. Sacher, "Inference for Regression with Variables Generated by AI or Machine Learning", arXiv:2402.15585 [econ.EM].

Bluhm, R., and M. Krause (2022), "Top lights: Bright cities and their contribution to economic development," *Journal of Development Economics* 157, 102880.

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, Fang H, Chatterjee S, 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.

Donaldson, D. and A. Storeygard (2016) "The View from Above: Applications of Satellite Data in Economics", *Journal of Economic Perspectives* 30, 171-198.

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

Frisch, R. (1934). Statistical confluence analysis by means of complete regression systems. University Institute of Economics, Oslo.

Gibson, J., O. Susan, and G. Boe-Gibson (2021), "Which night lights data should we use in economics, and where?" *Journal of Development Economics* 149, 102602.

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.

Li, X., Zhou, Y., M. Zhao and X. Zhao (2020), "A harmonized global nighttime light dataset 1992-2018". Scientific Data 7, 168. https://doi.org/10.1038/s41597-020-0510-y.

Lowes, S., and E. Montero (2021) "Concessions, violence, and indirect rule: evidence from the Congo Free State," *The Quarterly Journal of Economics* 136, 2047–2091.

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.

Schennach, S.M. (2016), "Recent Advances in the Measurement Error Literature", *Annual Reviews* 8, 341–377.

Storeygard, A. (2016) "Farther on down the road: Transport costs, trade, and urban growth in Sub-Saharan Africa," *The Review of Economic Studies 83*, 1263–1295.

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).

Young, A. (2013), "Inequality, the urban-rural gap, and migration", *The Quarterly Journal of Economics* 4, 1727–1786.