Empirical Economics (2023) 65:1215–1252
https://doi.org/10.1007/s00181-023-02377-y

Lights and GDP relationship: What does the computer tell us?

Diep Hoang Phan[1]

Received: 28 June 2022 / Accepted: 23 January 2023 / Published online: 9 February 2023

Abstract
The relationship between nighttime lights and GDP varies from country to country. However, which factors drive variations in the lights–GDP relationship across countries remains unclear. This paper examines the significance of approximately 600 potential drivers of uncertainty in the relationship between night lights and GDP worldwide. I employ three novel modern statistical techniques to select variables within a high-dimensional context: LASSO, minimax concave penalty, and spike-and-slab regression. Institutional quality emerges as the most important factor in explaining the difference between luminosity data and GDP across countries.

Keywords Nighttime lights · Modern statistical tools · Official GDP · Measurement errors

JEL Classification E01 · O47 · O11 · R14

1 Introduction

Gross domestic product (GDP) holds a crucial place in the social sciences and is a guiding principle for political decisions. Nevertheless, GDP is inadequately measured worldwide (Wu et al. 2013; Feige and Urban 2008). Reliable national GDP data are unavailable in many low- and middle-income countries due to statistical capacity and budget constraints (Keola et al. 2015). Local governments are likely to inflate real data in dictatorship nations, resulting in inadequate statistical data. Even high-income

I would like to extend my sincere thanks to the Editor, Joakim Westerlund, and a Reviewer for their constructive comments, which helped improve the paper. I gratefully acknowledge helpful discussion with Mita Bhattacharya, Russell Smyth, and Xibin Zhang and comments on earlier drafts from Paul Raschky, Weijia Li, Pushkar Maitra, and Sascha Becker.

✉ Diep Hoang Phan
diep.phan@monash.edu

[1] Department of Economics, Monash University, Caulfield East, VIC, Australia

countries suffer from measurement errors because they ignore the informal economy. Hence, dealing with measurement errors in GDP has stimulated economic research for many decades.

Recently, the absence of high-quality GDP data at the national and regional levels has forced many economists to use an alternative measure of regional outputs: nighttime lights (NTL). Luminosity or NTL can be detected by satellites from outer space. Exogenous characteristics, high spatial resolution, high-frequency accessibility, consistent quality, and global coverage are some of the key benefits of these data that make them appealing as an alternative measure of real GDP at various levels of subnational administrative areas. Therefore, luminosity is widely highlighted in the economic literature, especially in serving as an additional proxy for local economic outcomes (Martinez 2022; Hu and Yao 2021; Asher et al. 2021; Gibson et al. 2021; Chen and Nordhaus 2019; Keola et al. 2015; Hodler and Raschky 2014; Henderson et al. 2012; Chen and Nordhaus 2011). However, in contrast to the growing popularity of night lights in economic literature, our understanding of the main drivers of the uncertainty in the lights–GDP relationship remains unclear. Both NTL and GDP are subject to measurement errors, leading to erroneous results when their relationship is estimated. Therefore, understanding the hidden components of measurement error in assessing the lights–GDP relationship is a major concern for economists.

Economists attempt to cope with measurement errors. Table 1 presents details of several relevant studies dealing with measurement errors in the lights–GDP relationship. The existing literature focuses on two directions: (1) establishing a statistical framework to estimate errors, and (2) identifying specific elements that cause the difference between data observed from space and official measures of GDP. Although these approaches were practical, they failed to answer why the GDP and NTL relationship differ from country to country. The major limitations of the existing literature are their concentration on only a single or few factors that determine the variation between night lights and GDP (listed in Table 1). It is obvious that a country's GDP does not solely determine the amount of light consumed by its residents. If we consider NTL to be normal goods similar to other goods discussed in economics, consumer preferences will significantly influence their demand.[1] Thousands of factors may contribute to different NTL consumption preferences across countries. Therefore, we face a large number of potential drivers contributing to uncertainty in lights-GDP relationships with a relatively limited number of observations.[2] As a result, it is often difficult to clarify what eventually drives the difference between lights and GDP. For example, South Korea and Russia have similar GDPs, but differ greatly in population density, democracy level,[3] and share of the agricultural sector in GDP. If we focus on only one of the three factors listed above, it will be difficult to ascertain which one is the main factor affecting light consumption in each country. Even if attention is given to all three aspects simultaneously, there is a high probability of missing many other relevant factors that cause differences in the lights–GDP relationship in these

[1] In economic theory, normal goods are those whose demand rises when income increases and decreases when income falls, given that the price remains the same.

[2] This is because we only have 179 countries compared to thousands of factors.

[3] According to Freedom House, South Korea is classified as a democratic country while Russia is an autocratic country

two countries. Therefore, it is necessary to employ a high-dimensional approach that considers thousands of elements simultaneously.

Accordingly, to optimize GDP estimation using NTL at national and subnational levels, economists and researchers must better understand the influencing factors of the lights–GDP relationship. This study aims to identify the factors that determine the variation between NLT and GDP across countries. In particular, this paper employs three modern statistical tools: the least absolute shrinkage and selection operator (LASSO), the minimax concave penalty, and the spike-and-slab regression, to examine a dataset of approximately 600 potential drivers. There are several advantages of this approach in selecting essential predictors, including the following:

1) These methods have the advantage of considering all potential factors but only selecting a subset of covariates.
2) These methods allow the computer to automatically choose important regressors without the bias of the researcher's subjective view.
3) Modern statistical tools such as LASSO can evaluate the relative importance of each factor. This application is especially significant, as the relative importance of the regressors is often the primary motivation for analyzing the lights–GDP uncertainty.
4) Since the three methods are based on different algorithms and theories, I also seek to prove that the findings are robust and do not omit any critical factors.

The results indicate that the quality of the institution is the main factor that determines the variation between NTL and GDP across countries. The author found multiple indicators reflecting institutional quality, ranging from the degree of democracy to the number of years the leader has spent in the office and the government's effectiveness at controlling corruption and resolving conflicts. In addition, the business environment and the level of development are also important factors. Furthermore, other factors such as economic structure, urbanization, and geography also significantly affect the lights–GDP relationship.

The remainder of this paper is organized as follows: Sect. 2 describes the theoretical framework. Section 3 summarizes the variable selection methods used in this paper. Section 4 provides a brief overview of the dataset used at the national level. Section 5 presents the empirical findings of this study, while Sect. 6 discusses the findings of this study. Finally, Sect. 7 concludes and highlights the potential for future research.

2 Theoretical framework

Many studies have shown that aggregate lights per area are positively correlated with GDP in that area. Doll et al. (2000) used the log-log model to examine the linear relationship between the purchasing power parity (PPP) GDP and total lit area worldwide for 1994–1995, obtaining an R-square value of 0.85. Ghosh et al. (2010) derived an R-square value of 0.73 by regressing PPP GDP and the total amount of lights worldwide in 2006.

Table 1 Selected studies on dealing with measurement errors in lights–GDP estimation

No.	Study	Method/approach	Contribution	Limitations
1	Henderson et al. (2012)	Classical measurement error model		
2	Chen and Nordhaus (2011)	Grouped countries into five categories based on statistical capacity to estimate the optimal weights for luminosity and GDP to reduce the estimation errors	Statistical framework	Overdependence on luminosity data and, therefore, tendency to overestimate or underestimate the components of GDP that emit too much or little light
3	Hu and Yao (2021)	New nonclassical measurement error model		
4	Ghosh et al. (2010)	Identifying the informal economy as part of errors in the lights–GDP relationship		Only focus on one or a few factors affecting the relationship between NTL and GDP
5	Wu et al. (2013)	Adding GDP per capita, latitude, spatial distribution of human activities and gross savings rate into the estimation of NTL and GDP	Identifying factors affecting the lights–GDP relationship	Depends on the subjective view of the researcher in selecting factors (or variables)
6	Keola et al. (2015)	Adding the agricultural sector into the estimation of NTL and GDP		Due to consideration of a single factor,. it cannot evaluate each factor's relative importance. As a result, we still cannot clarify what eventually drives the difference between lights and GDP
7	Martinez (2022)	Adding the degree of democracy into the estimation of NTL and GDP		Thousands of other factors have still not been examined yet

Henderson et al. (2012) suggested the following equation:

$$\text{light/area} = \phi(\text{GDP/area}) = \beta(\text{GDP/area})^{\alpha}. \tag{1}$$

This paper is based on Wu et al. (2013)'s article. Our study differs from that of Wu et al. (2013) in that it simultaneously analyzes 600 dimensions that might affect the lights–GDP relationship instead of considering only three factors. Furthermore, our approach allows the computer to automatically select factors without the bias of the researcher's subjective view.

Wu et al. (2013) hypothesized that the amount of lights is a power function of the GDP in each nation:

$$\text{light} = \phi(\text{GDP}) = k(\text{GDP})^{\alpha}, \tag{2}$$

where parameter k is not a constant, and a number of unknown factors other than GDP identify it. The hidden components of k are the main focus of this paper. Several factors might be potential elements of k, for example, income per capita. This is because a higher income per capita level definitely increases the consumption of normal goods, including lights. The share of the agricultural sector would be another possible element, as a higher portion of agriculture in the GDP often results in a lower light demand for residences at night. Another aspect to consider is population density. For example, while Russia and South Korea have similar GDPs, their light intensities differ significantly, which may result in different light consumption. Since there are hundreds of potential factors, we still do not know which factors significantly affect parameter k.

Therefore, the parameter k can be decomposed and allocated to several variables:

$$k = k_0 e^{k_1 x_1} e^{k_2 x_2} e^{k_2 x_3} \cdots e^{k_n x_n}, \tag{3}$$

where k_0 is constant, $x_1, x_2, \ldots, x_n$ are unknown factors, and k_1, k_2, …, k_n are the respective coefficients of the variables above. Taking the logarithmic transformation, we obtain the following:

$$\ln(\text{light}) = \ln(k_0) + \alpha \ln(\text{GDP}) + k_1 x_1 + k_2 x_2 + \cdots + k_n x_n, \tag{4}$$

Since different satellites or the same satellites in different years obtain different images, they cannot be directly compared. To eliminate these obstacles, we introduce time dummies into the model.

$$\ln(\text{light}_{it}) = \delta_t + \ln(k_0) + \alpha \ln(\text{GDP}_{it}) + \eta_{it}, \tag{5}$$

where i indexes the country, t indexes the year, δ_t is the time dummy, and

$$\eta_{it} = k_1(x_1)_{it} + k_2(x_2)_{it} + \cdots + k_n(x_n)_{it} + \varepsilon_{it}, \tag{6}$$

where ε_{it} is a random error term.

To control factors that vary from country to country, Henderson et al. (2012) used country-fixed effects. As this paper examines these factors, I do not adopt country-fixed effects. Instead, I use the absolute mean value of $\widehat{\eta}_{it}$ obtained from (5) as the dependent variable. On the right-hand side of the equation, I test approximately 600 variables representing several aspects of a country (including time-variant[4] and time-invariant variables) since we do not know which specific elements significantly affect the parameter k. These variables include the quality of the political institution, degree of democracy, economic structure, geography, demographics, infrastructure, urbanization, energy consumption, natural resources, foreign aid, remittances, statistical capacity score, cultural diversity, religion, history, lands, and climate, among others. I employ modern statistical tools, including LASSO, spike and slabs, and the minimax concave penalty, to select the most important predictors. Therefore, (6) becomes[5] the following:

$$|\bar{\widehat{\eta}}_i| = \gamma \boldsymbol{X_i} + \mu_i, \tag{7}$$

where $|\bar{\widehat{\eta}}_i|$ is the absolute mean value of error terms (grouped by each country) obtained from (5), and $\boldsymbol{X_i}$ is a set of control variables (approximately 600 variables). The next step is to perform regressions using (7).

3 Variable selection methods

We are confronted with the problem of a high-dimensional data context. While the dataset has only 179 observations of the dependent variable corresponding to 179 countries globally, thousands of explanatory variables may significantly affect the lights–GDP relationship across countries. To address the high-dimensional nature of the data and ensure the objectivity of the results, this paper utilizes three methods of variable selection called modern statistical techniques. Specifically, I used three alternative methods to choose variables: LASSO, the minimax concave penalty, and spike-and-slab.

3.1 LASSO

The LASSO model works effectively with relatively many predictors and a low number of observations. This technique is based on the shrinkage of the least-squares regression coefficients. This process leads to some parameter estimates being set to precisely zero. In other words, the purpose of LASSO is to eliminate useless variables from the model and retain only the most important independent variables in explaining the outcome variable. This strategy allows the variable selection to be automated with high

[4] For time series variables, we obtain the average for the values throughout the study period.

[5] I average the error terms from Eq. 5 to ensure that it is applicable to use variable selection methods. For example, LASSO requires that the number of observations should be less than the number of predictors. I cannot meet this condition if I use panel data. This approach will not affect the findings since several existing studies found that the lights–GDP relationship does not change much over time but across countries. Alternatively, I can select a specific year to perform the analysis. However, this might result in bias in our conclusion.

accuracy. Another advantage of LASSO is that it is computationally efficient [see for example,Varian 2014].

LASSO was first proposed by Tibshirani (1996). This method was presented in detail in Bühlmann and Van De Geer (2011) (page 7–43). It is challenging to model high-dimensional data. For a continuous response variable $Y \in \mathbb{R}$, the linear model is a simple yet very useful solution:

$$Y_i = \sum_{j=1}^{p} \beta_j X_i^{(j)} + \varepsilon_i, \tag{8}$$

for $i, \ldots, n$, where $\varepsilon_1, \varepsilon_2, \varepsilon_3, \ldots, \varepsilon_n$ are independent and identically distributed (iid) and independent of X_i, and it is assumed that $E[\varepsilon_i] = 0$.

The matrix- and vector-notation form of (8) is:

$$Y = X\beta + \varepsilon,$$

with the response vector being represented by Y_{n+1}, the design matrix by $X_{n\times p}$, the parameter vector by $\beta_{p\times 1}$, and the error vector by $\varepsilon_{n\times 1}$.

The ordinary least-squares estimator is not unique when $p > n$ and substantially overfits the data. Therefore, complexity regularization is necessary. Here, we use regularization with the $\ell 1$-penalty. LASSO is used to estimate the parameters in model (8):

$$\hat{\beta}\,(\lambda) = \arg\min_{\beta} \left(\frac{\|Y - X\beta\|_2^2}{n} + \lambda\|\beta\|_1 \right), \tag{9}$$

where $\|Y - X\beta\|_2^2 = \sum_{i=1}^{n} (Y_i - (X\beta)_i)^2$, and $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$. In the above equation, $\lambda \geq 0$ is a tuning parameter controlling the power of the penalty, and a larger λ corresponds to a larger shrinkage of the model. $\lambda = 0$ indicates that the problem becomes the ordinary least-squares fit. When $\lambda = \infty$ or as λ becomes sufficiently large, it indicates that all parameter estimates are forced to be zero.

The estimator performs variable selection in the sense that $\hat{\beta}_j(\lambda) = 0$ for some j's (depending on the choice of λ), and $\hat{\beta}_j(\lambda)$ can be considered a shrunken least-squares estimator. This results in the exclusion of features with coefficients from the model equal to zero. LASSO is therefore a powerful method for selecting features.

3.2 The minimax concave penalty (MCP)

The MCP can yield nearly unbiased shrinkage estimates as a possible alternative to the LASSO penalization method. In particular, Zhang (2010) examined the properties of the MCP for linear regression in a high-dimensional context and found that it provides continuous, nearly unbiased, and accurate variable selection. In this section, I will provide a brief description of MCP (Breheny 2016). In the literature, one can find more detailed discussion about MCP (see for example,Zhang 2010; Breheny and Huang 2011).

Let us consider a regression analysis with response $y \in \mathbb{R}^n$ and design matrix $X \in \mathbb{R}^{n\times p}$. The MCP is an alternative method used to obtain more accurate regression coefficients in sparse models. This technique was first introduced by Zhang (2010) by considering the objective function:

$$Q(\beta|X, y) = \frac{1}{2n}\|y - X\beta\|^2 + \sum_{j=1}^{p} P(\beta_j|\lambda, \gamma), \tag{10}$$

where $P(\beta|\lambda, \gamma)$ is a folded concave penalty.

Unlike LASSO, many concave penalties depend on λ in a non-multiplicative way, so that $P(\beta|\lambda) \neq \lambda P(\beta)$. In addition, they typically involve a turning parameter γ that controls the concavity of the penalty.

The formula behind the MCP is expressed as follows:

$$P_\gamma(x;\lambda) = \begin{cases} \lambda|x| - \frac{x^2}{2\gamma}, & \text{if} \quad |x| \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & \text{if} \quad |x| > \gamma\lambda \end{cases}, \tag{11}$$

For $\gamma > 1$. Its derivative is

$$\dot{P}_\gamma(x;\lambda) = \begin{cases} \left(\lambda - \frac{|x|}{\gamma}\right) sign(x), & \text{if} \quad |x| \leq \gamma\lambda \\ 0, & \text{if} \quad |x| > \gamma\lambda \end{cases}. \tag{12}$$

MCP starts by applying the same rate of penalization as LASSO and then smoothly relaxes the rate to zero as the absolute value of the coefficient increases.

Among all penalty functions that are continuously differentiable on $(0, \infty)$ and satisfy $\dot{P}(0+;\lambda) = \lambda$ and $\dot{P}(t;\lambda) = 0$ for all $t \geq \gamma\lambda$, MCP minimizes the maximum concavity as follows:

$$\kappa = \underbrace{\sup}_{0<t_1<t_2} \frac{\dot{P}(t_1;\lambda) - \dot{P}(t_2;\lambda)}{t_2 - t_1}. \tag{13}$$

3.3 Spike-and-slab

A Bayesian technique to choose variables called spike-and-slab regression is a novel approach for economists. This method is described in detail in Ishwaran and Rao (2005). This section will present a brief introduction to spike-and-slab regression (see Varian 2014).

We consider a linear model with P possible predictors. Then, γ is denoted as a vector of P-dimensional consisting of zeros and ones, indicating whether a particular variable appears in the regression.

In the first step, a Bernoulli prior distribution is applied to γ; for example, we might initially assume all variables have a similar probability of being included in the regression. Then, conditional on a variable being in the regression, we define a prior distribution as per its regression coefficient. For example, we might use a normal prior with a mean of 0 and a large variance. The method's name comes from these two priors: the "spike" is the probability that a coefficient will be nonzero; and the "slab" is the (diffuse) prior that describes the possible values for the coefficient.

The next step is to sample γ from its prior distribution. This will result in a set of variables used in the regression. Based on this list of included variables, we draw coefficients from the prior distribution. By combining the two draws with the likelihood, we obtain a posterior distribution on the probability of inclusion and the coefficients. Through Markov Chain Monte Carlo (MCMC) simulation, we repeat this process thousands of times, giving us a summary table of the posterior distribution for γ (including variables), β (the coefficients), and the predictions associated with the prediction of y. There are various ways to summarize this table. For example, by computing the average value of γp, we can demonstrate the posterior probability of the variable p appearing in the regressions.

4 Data

4.1 Nighttime lights and GDP data

In this paper, I use data from the Defense Meteorological Satellite Program (DMSP) as the primary source for measures of NTL and GDP is calculated from the replication files of Pinkovskiy and Sala-i Martin (2016).[6] These replication data guarantee that the results below are not affected by ad hoc selections regarding variables and data sources (Martinez 2022). Furthermore, the results below are comparable to many key studies in the literature(Martinez 2022; Pinkovskiy and Sala-i Martin 2016; Keola et al. 2015; Chen and Nordhaus 2011). This dataset covers the period from 1992 to 2010, and Pinkovskiy and Sala-i Martin (2016) used "GDP per capita, PPP, constant 2005 international dollars" from the World Bank's World Development Indicator (WDI). This variable contains data from almost all countries without missing values.

Observations of light at night are collected, processed, and maintained by the National Oceanic and Atmospheric Administration (NOAA). Nighttime luminosity is available at the pixel-year level (approximately 0.86 square kilometers at the equator) from 1992 to 2013. The intensity of lights is represented by a six-bit digital number (DN) in a grid format. Digital numbers range from 0 (no light) to 63 (top-coded). Adding all the digital numbers across pixels produces a light proxy for aggregate income:

$$\text{Light}_{j,t} = \sum_{i=1}^{63} i * (\#\,\text{of pixels in country } j \text{ and year } t \text{ with DN} = i).$$

According to Henderson et al. (2012), the logarithms of the aggregate luminosity measure will be averaged when there are multiple satellite measurements in a given year. The literature widely uses this formula as a standard practice (Chen and Nordhaus 2011; Henderson et al. 2012; Martinez 2022).

[6] I would like to thank Dr. Maxim Pinkovskiy for providing us with data on NTL.

With DMSP NTL data from 1992 to 2010, various concerns related to blurring, top-coding, and lack of calibration (Gibson et al. 2021) arise. Therefore, I will conduct various robustness checks with newer and better lights and GDP data to address this problem. In particular, I use a harmonized global NTL dataset from 1992 to 2018, a newer NTL dataset with a longer period. This dataset is obtained in GeoTIFF file format from the open-source database Scientific Data published by *Nature*[7] (Li et al. 2020).

The harmonized dataset is globally integrated and consistent, combining the inter-calibrated NTL observations from the DMSP data with the simulated DMSP-like NTL observations from the VIIRS data. The global DMSP NTL time series (1992–2018) reveals consistent temporal trends. There is no separate quality file since the data are already produced with quality weights. I downloaded and processed the GeoTIFF file with R software for a global scale. Corresponding with alternative NTL data, I also used a newer vintage of GDP data—GDP per capita, PPP, and constant 2017 international dollars. Figure 1 presents scatter plots of log lights per capita (or log aggregate lights per area) against log GDP per capita using two alternative sources of NTL and GDP data.

4.2 Other data

The rest of the analysis variables come from various data sources, including the World Development Indicator (WDI),[8] Freedom House,[9] Quality of Government (QoG),[10] Varieties of Democracy dataset,[11] WHOGOV dataset (Nyrup and Bramwell 2020),[12] Center of Systemic Peace (Marshall et al. 2011),[13] and others. These variables describe the quality of the political institution, degree of democracy, economic structure, geography, demographics, infrastructure, urbanization, energy consumption, natural resources, foreign aid, remittances, statistical capacity score, cultural diversity, religion, history, lands, and climate, among others. For example, to evaluate the effect of institutional quality on the lights-output relation, I intentionally use the Freedom in the World (FiW) index to ensure that the results below are comparable to the work of Martinez (2022). These data are published by Freedom House annually. Freedom House divides countries into three groups: “free, ”“partially free,” and “not free.” In this paper, instead of using the FiW index as a time series variable, I use it as a cross-sectional variable[14] to help uncover the relationship of political regimes with the difference between lights and GDP. Table 2 provides an overview of these data.

[7] The data are available at: https://figshare.com/articles/dataset/Harmonization_of_DMSP_and_VIIRS_nighttime_light_data_from_1992-2018_at_the_global_scale/9828827.

[8] Available at: https://databank.worldbank.org/source/world-development-indicators.

[9] Available at: https://freedomhouse.org/.

[10] Available at: https://www.gu.se/en/quality-government/qog-data/data-downloads.

[11] Available at: https://www.v-dem.net/vdemds.html.

[12] Available at: https://politicscentre.nuffield.ox.ac.uk/whogov-dataset/.

[13] Available at: https://www.systemicpeace.org/.

[14] This is because as a category variable, the freedom status of a country does not change much over a short period.

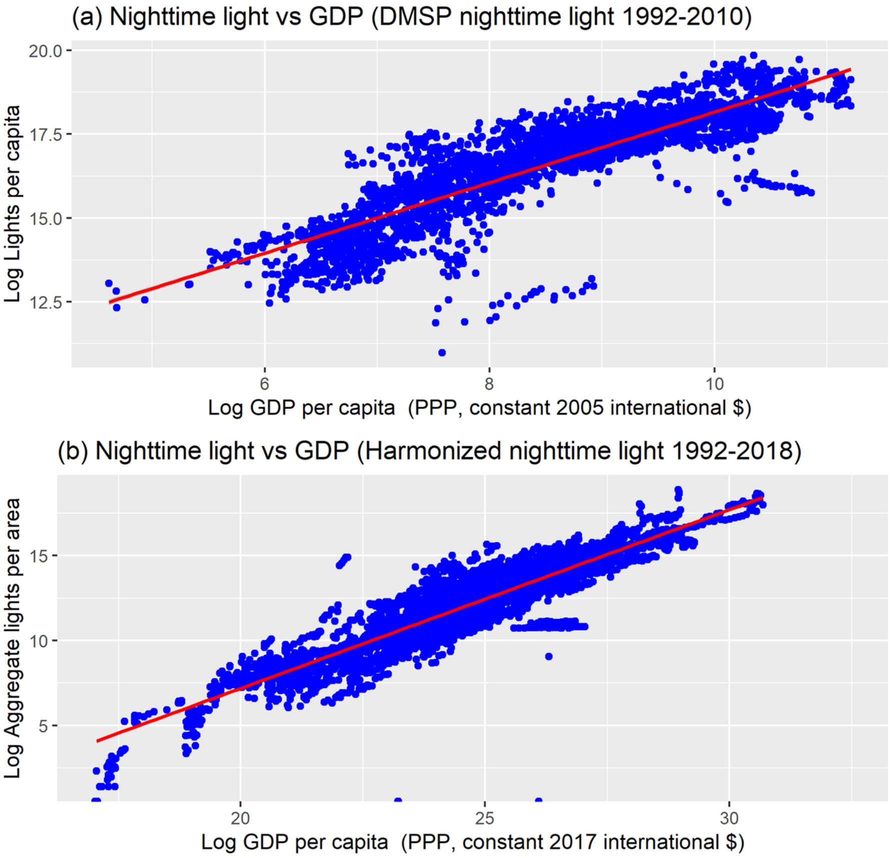

Fig. 1 Official GDP and lights

Table 2 Distribution of countries by freedom status in 2010. *Source*: Freedom House

Freedom status	Number of countries	Representative countries
Free	81	Australia, Canada, USA
Partially free	56	Albania, Sri Lanka, Philippines
Not free	41	Vietnam, China, Russia
Total	178	

For data sources and definitions of all key variables in this paper, please refer to Appendix B and Appendix C. Table 6 in Appendix A shows the summary statistics for some key variables in this paper.

5 Results

Figure 2 shows a scatterplot of log real GDP (PPP) per capita (2005 US dollars) for 2010 and the absolute value of error terms ($N = 133$). There is a significant difference between the size of error terms across countries. On the one hand, although the UK, Germany, and France are located in the same geographical region (Western Europe) and have similar GDP per capita, the absolute values of the error terms are very different. We can also see similar patterns in some Southeast Asian countries (Philippines, Indonesia, and Vietnam) and sub-Saharan countries (Kenya, Ghana, and Lesotho). On the other hand, India and France clearly come from different income groups and geographical regions, but the magnitudes of their error terms are the same. This paper explores the kind of unobservable information contained in error terms that help us understand the difference between lights and official reported GDP across countries.

Table 3 illustrates the results of three alternative variable selection methods (I present the results in detail in Online Appendix D). First, I examine a dataset of 597 variables and 172 countries to explore the factors that determine the discrepancy between lights and GDP. The table presents 12 variables selected by LASSO, spike-and-slab, and MCP regressions. Digits in each column represent the ordinal importance of the variable, and dashes indicate that a variable was excluded from the chosen model (I excluded all other irrelevant variables). Table 3 highlights two key facts. First, the three methods draw consistent results. In other words, they selected similar variables. Second, most of the variables reflect the political institution's quality. On the one hand, many factors determine the degree of autocracy, including freedom status (Martinez 2022), regime type, and the consecutive number of years the leader has been in office (Nyrup and Bramwell 2020). On the other hand, other variables such as starting a business score, state fragility index, public sector corruption index, or natural resource protection indicator measure a government's effectiveness. In addition, the statistical techniques also identify other elements that might significantly affect error terms, such as geography (average distance to nearest ice-free coast) and level of economic development (agriculture, forestry, fishing, value-added). Finally, I move to the subsequent analysis to see how well these modern statistical techniques perform in selecting important predictors of error terms.

To conduct cross-sectional analysis, I estimate Eq. (7):

$$|\hat{\bar{\eta}}_i| = \gamma \boldsymbol{X_i} + \mu_i.$$

Tables E2–E12 in online Appendix E report the simple linear regression of the absolute mean value of error terms on various control variables. In this analysis, in addition to using variables selected by three alternative statistical methods in Table 3, I also controlled for a diverse group of variables representing political, geographical, economic development, ethnic and cultural diversity, land, and historical and demographic factors. This is to ensure I do not omit any essential predictors in controlling for discrepancies across countries and for comparison purposes. Figure 3 plots all point estimates of all variables I tested in Tables E2–E12. Note that we use standardized variables; thus, the coefficients can be comparable. As we can see from Fig. 3, all variables that the LASSO, spike-and-slab and the MCP regressions select (in Table 3) have a relatively large effect on the absolute mean value of error terms (see red line in Fig. 3). In contrast, all other variables (which the three above models do not choose) are statistically nonsignificant (blue line) or statistically significant but economically nonsignificant (yellow line) except for the service sector as a share of GDP variables and urban population growth rate. The negative signs on the coefficients of the agricultural and service sectors indicate that the development level considerably affects the relationship between lights and GDP. Specifically, while lights typically more accurately predict GDP for countries with higher service sector shares (a negative sign), they are worse for nations with a high percentage of the agricultural sector (a positive sign). Additionally, in univariate linear regression results, the sign of all coefficients is consistent with the three statistical models in Table 3) as well as our expectations (for details see online Appendix E). The results suggest that all variable selection methods perform fairly well.

Table 4 describes the multivariate analyses. Multiple regression generally confirmed the results of the simple linear regression. However, the more variables that indicate the quality of an institution, the higher the chance of collinearity. As a result, I divided these variables and controls into separate regressions. It is clear that the coefficients of the univariate linear regression and the multivariate linear regression are generally of a similar magnitude and sign across all variables (see the same variables in Table 4 and Tables E2 and E3 in online Appendix E). It should be emphasized that coefficients on individual variables in Table 4 generally follow the expected direction. For example, the sign (positive) and magnitudes of the coefficients for the average distance to the coast are consistent and stable across regressions (see row 10 Table 4 and column 2 Table E3 in online Appendix E). Therefore, I expect that lights will be a better proxy for GDP if a country is located close to the coast. Another example shows that the absolute mean value of the error term for non-free nations will be significantly higher than that for partly free and free countries (see column 1). Additionally, the consecutive number of years the leader has spent in office also reflects the status of the degree of democracy. Columns 3 and 8 show that the signs of the coefficient of this variable are positive. In many dictatorships, leaders have been in positions for many years. Thus, autocracy regimes may manipulate GDP. This finding provides additional evidence for the conclusions drawn by Martinez (2022). In his research, he concluded, *"I estimate that the most authoritarian regimes inflate yearly GDP growth rates by a factor of 1.15–1.3 on average"* (page 28).

In conclusion, this section presents the results of the cross-sectional analysis. With the help of three alternative variable selection methods, I systematically analyze hundreds of variables. The results show that the degree of democracy, government effectiveness, and level of development are the key determinants of the discrepancy between lights and GDP. In addition, distance to the coast and urban population growth rate also significantly impact the lights–GDP association.

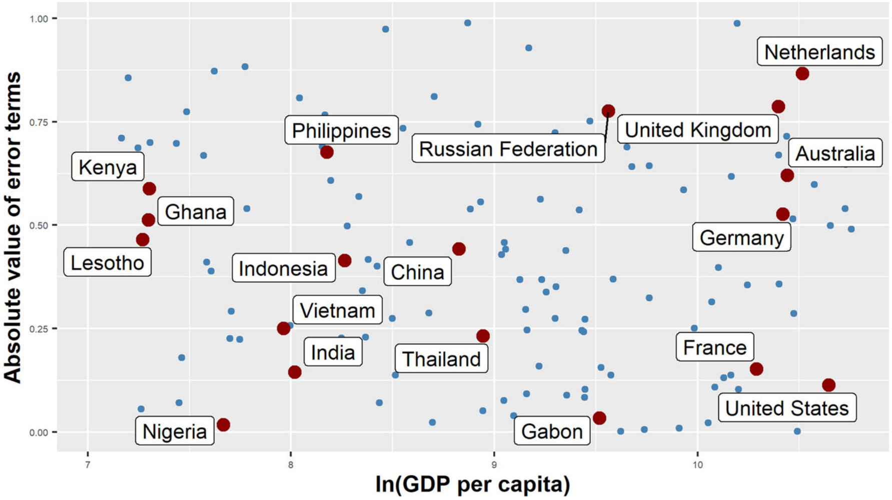

Fig. 2 GDP per capita and error terms, all countries. *Notes* The figure shows the scatter plot of log real GDP (PPP) per capita for 2010 and the absolute value of the error term. The error terms come from the regression of log light per capita on log GDP per capita with year fixed effects. Additionally, I highlighted some countries in dark red. Blue points represent all other countries. ($N = 133$) (Color figure online)

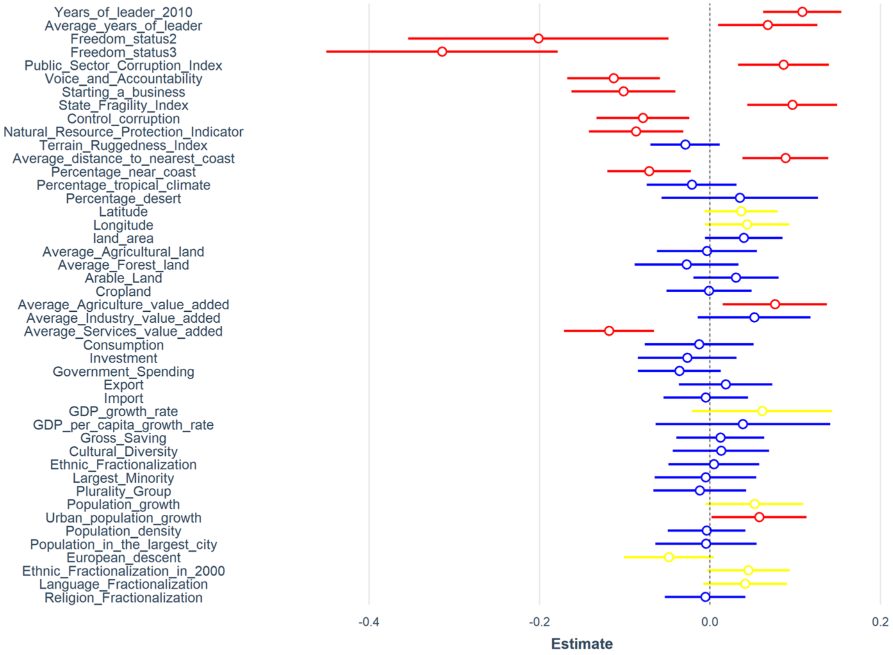

Fig. 3 Multiplot coefficients of control variables from Eq. (7). *Notes* I use DMSP NTL and GDP per capita, PPP (in constant 2005 international $) (1992–2010). This figure summarizes the result from Tables E2–E12 in online Appendix E. In this figure, I already standardized all control variables; thus, all coefficients are comparable. (Note that in Tables E2–E12, I report the value of coefficients of control variables without standardization)

Table 3 Comparison of variable selection methods: which factors determine the difference between lights and GDP?

Predictor	LASSO	Spike-and-slab	Minimax concave penalty
Freedom status	1	1	–
Perception of Electoral Integrity Index	2	17	10
State Fragility Index	3	4	–
Starting a business score	4	3	2
Parliamentary election: compulsory voting	5	16	6
Natural resource protection indicator	6	6	7
Regime type	7	14	5
Average distance to nearest ice-free coast (1000 km.)	8	13	8
Number of years the leader in office continuously in 2010	9	2	1
Public sector corruption index	10	5	3
Agriculture, forestry, and fishing, value added (% of GDP)	11	11	9
Voice and accountability, estimate	12	7	–

Table 3 illustrates the results of different methods in selecting variables. I examine a dataset of 597 variables and 172 countries to explore which factors determine the discrepancy between lights and GDP. The table presents twelve variables that were selected by LASSO, spike-and-slab, and minimax concave penalty regressions. Digits in each column represent the ordinal importance of the variable, and dashes indicate that a variable was excluded from the chosen model

Table 4 Cross-sectional analysis: multiple linear regression

	Dependent variable							
	Absolute mean value of error terms (1992–2010)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Freedom status 2	− 0.1503**							
	(0.0699)							
Freedom status 3	− 0.2400***							
	(0.0672)							
Average starting a business score		− 0.0047***						
		(0.0017)						
Years leader is in office (2010)			0.0112***					
			(0.0026)					
Average public sector corruption index				0.2471**				
				(0.0968)				
Average voice and accountability					− 0.0886***			
					(0.0329)			
Average State Fragility Index						0.0131**		
						(0.0052)		
Natural resource protection indicator							− 0.0022***	
							(0.0008)	

Table 4 continued

	Dependent variable							
	Absolute mean value of error terms (1992–2010)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average years of leader in office								0.0091**
								(0.0044)
Average distance to coast	0.1585**	0.2165***	0.1935***	0.1870**	0.1669**	0.1795**	0.1941***	0.2233***
	(0.0708)	(0.0774)	(0.0735)	(0.0754)	(0.0728)	(0.0790)	(0.0729)	(0.0753)
Ethnic fractionalization (2000)	0.0074	− 0.0726	0.0050	− 0.0231	− 0.0290	− 0.0256	0.0465	0.0182
	(0.1488)	(0.1492)	(0.1511)	(0.1486)	(0.1546)	(0.1516)	(0.1405)	(0.1532)
Language fractionalization (2000)	0.0219	0.0098	0.0508	0.0177	0.0179	− 0.0650	0.0573	0.0274
	(0.1301)	(0.1281)	(0.1309)	(0.1320)	(0.1323)	(0.1383)	(0.1249)	(0.1338)
Religion fractionalization (2000)	0.0132	0.0144	− 0.0014	− 0.0070	0.0082	0.0012	− 0.0307	− 0.0227
	(0.0968)	(0.0962)	(0.0926)	(0.0992)	(0.1003)	(0.0999)	(0.1005)	(0.0979)
Average GDP growth (annual %)	0.0111	0.0172*	0.0112	0.0152	0.0117	0.0152	0.0178	0.0153
	(0.0126)	(0.0103)	(0.0109)	(0.0118)	(0.0124)	(0.0122)	(0.0130)	(0.0113)
Population density	0.0001	0.0001	0.0001	0.0001	0.00005	0.0001	0.00002	0.0001
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Observations	172	172	172	172	172	172	172	172
R^2	0.1698	0.1588	0.1659	0.1416	0.1496	0.1417	0.1498	0.1285
Adjusted R^2	0.1291	0.1229	0.1303	0.1050	0.1133	0.1050	0.1135	0.0913

The table reports OLS coefficient estimates from a multiple linear regression of mean absolute errors (1992–2010) on the variables shown. Note that the three variable selection methods all select the first nine variables. Column 1 shows the coefficient of freedom status and other variables. Freedom status is a value draw from the Freedom in the World (FiW) index published annually by Freedom House. In this paper, I used the FiW index for 2010. There are three statuses: 1. Not free countries; 2. Partly free countries; 3. Free countries. For other time series variables, I used an average value for 1992–2010. I also employ predictive mean matching to impute missing values (the rate of missing data is shallow). The authors downloaded data from various sources: the WhoGov dataset, Freedom House, Varieties of Democracy dataset, Center of Systemic Peace, World Development Indicator, and Quality of Government. Robust standard errors are in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

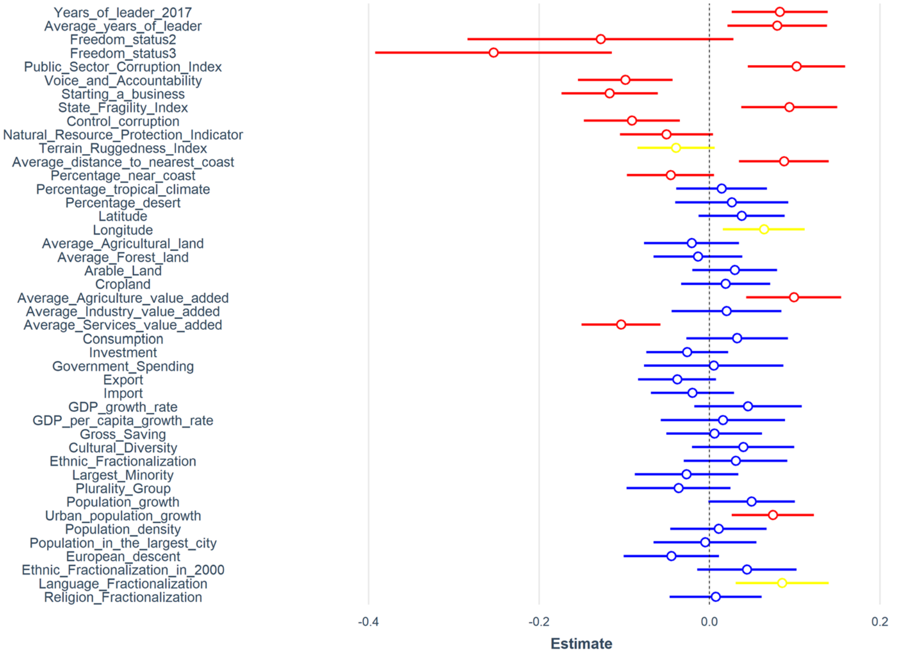

Fig. 4 Multiplot coefficients of control variables from Eq. (7). *Notes* I use harmonized NTL and GDP per capita, PPP (in constant 2017 international $) (1992–2018). This figure summarizes the result from Tables F2–F12 in online Appendix F. In this figure, we already standardized all control variables; thus, all coefficients are comparable. (Note that in Tables F2–F12, we report the value of coefficients of control variables without standardization)

5.1 Robustness check

The previous sections use DMSP NTL data and GDP per capita, PPP (in constant 2005 international $) replicated from the replicate file from data used by of Pinkovskiy and Sala-i Martin (2016) to examine the factors determining the difference between lights and GDP. Our primary purpose is to compare our significant findings with certain popular papers in the literature, such as Martinez (2022), Pinkovskiy and Sala-i Martin (2016), Keola et al. (2015), Henderson et al. (2012) and Chen and Nordhaus (2011). However, there are two limitations to this dataset. First, the data are slightly outdated, with a limited time frame between 1992 and 2010. Second, DMSP NTL is affected by various flaws, such as blurring, coarse resolution, no calibration, low dynamic range, and top-coding (Gibson et al. 2021). Therefore, this section uses alternative NTL and GDP data to check the robustness of our findings. Specifically, I use harmonized NTL data, which are longer and better DSMP-like NTL data, from 1992 to 2018. In addition, I use a newer vintage of GDP data - GDP per capita, PPP (in constant 2017 international $).

The difference between lights and GDP is primarily a result of the differences in the quality of institutions. Therefore, I only focus on the replication of the cross-sectional analysis. Table 5 and Fig. 4 present the results of the cross-sectional analysis using new

NTL and GDP data. For all control variables, the sign and magnitude of all coefficients are similar (compare Figs. 3 and 4). In a similar vein, the multivariate analysis also draws consistent results (compare Tables 4 and 5). Therefore, the choice of whether to use the newer and longer NTL data should not be the main concern when assessing the relationship between lights and GDP. (See more details in online Appendix F).

6 Discussion

This paper examines the factors affecting the variation in the relationship between NTL and GDP across countries. I selected and processed a dataset of 600 potential drivers from various aspects, including institutional quality, degree of democracy, economic structure, geography, demographics, infrastructure, urbanization, energy consumption, natural resources, foreign aid, remittances, statistical capacity score, cultural diversity, religion, history, land, and climate, among others. I applied three modern statistical tools to select variables within a high-dimensional context: LASSO, MCP, and spike-and-slab regression. The results suggest that the cross-sectional discrepancy in the light-GDP relationship comes primarily from the quality of the institution. Our estimates show a high correlation between error terms and various indicators reflecting institutional quality. This includes the degree of democracy, the duration of a leader's tenure in office, the government's effectiveness in controlling corruption and its conflict resolution capabilities, the business environment, and development levels. In addition, geographic determinants such as average distance to the nearest ice-free coast considerably affect the lights–GDP relationship through benefits from trade (Henderson et al. 2012). Furthermore, urbanization is another influencing factor. It is also important to highlight that the growth rate in light might not capture the growth rate of the urban population in some regions. Our findings are robust when we use alternative NTL and GDP data.

The strong association between institutional quality and the discrepancy in the lights–GDP relationship remains a puzzle. One possibility is that many autocratic regimes manipulate GDP numbers (Martinez 2022). Additionally, our evidence indicates that the duration of the leader's tenure in office enhances inflated accounts of national statistics in dictatorships, causing a higher value in the error terms. Furthermore, the capacity to combat corruption significantly affects the measurement errors for standard output data. Finally, the level of development reflects statistical capacity.

7 Conclusion

In summary, the uncertain association between nighttime light data and national output is a major concern, particularly in proxy research. This study provides a broad picture of factors that identify the discrepancy between lights and GDP globally. One main assumption used widely as a standard practice in the literature is that the elasticity between lights and GDP is roughly constant across time and space (Henderson et al. 2012; Pinkovskiy and Sala-i Martin 2016). However, our findings suggest that the elasticity between luminosity data and official national accounts varies across time

Table 5 Cross-sectional analysis: multiple linear regression

	Dependent variable							
	Absolute mean value of error terms (1992–2018)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Freedom status 2	− 0.0759							
	(0.0726)							
Freedom status 3	− 0.2024***							
	(0.0706)							
Average starting a business score		− 0.0080***						
		(0.0022)						
Years of leader in office (2017)			0.0069**					
			(0.0027)					
Average public sector corruption index				0.3370***				
				(0.1095)				
Average voice and accountability					− 0.0877**			
					(0.0371)			
Average State Fragility Index						0.0112*		
						(0.0063)		
Natural resource protection indicator							− 0.0011	
							(0.0008)	
Average years of leader in office								0.0100**
								(0.0045)

Table 5 continued

	Dependent variable							
	Absolute mean value of error terms (1992–2018)							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Average distance to coast	0.1593**	0.2182***	0.1769**	0.1663**	0.1568**	0.1690**	0.1982***	0.2023***
	(0.0716)	(0.0757)	(0.0720)	(0.0715)	(0.0742)	(0.0792)	(0.0722)	(0.0708)
Ethnic fractionalization (2000)	−0.2537	−0.2862	−0.1692	−0.2724	−0.2482	−0.2441	−0.1642	−0.2120
	(0.1973)	(0.1984)	(0.2036)	(0.1878)	(0.1953)	(0.1976)	(0.1939)	(0.1990)
Language fractionalization (2000)	0.2752	0.2252	0.2677	0.2617	0.2810	0.2203	0.2905*	0.2701
	(0.1713)	(0.1724)	(0.1774)	(0.1674)	(0.1709)	(0.1805)	(0.1704)	(0.1738)
Religion fractionalization (2000)	0.0602	0.0393	0.0157	0.0354	0.0479	0.0299	0.0103	0.0320
	(0.1025)	(0.0974)	(0.1039)	(0.0988)	(0.1006)	(0.0999)	(0.1032)	(0.1002)
Average GDP growth (annual %)	0.0055	0.0055	0.0115	0.0051	0.0027	0.0078	0.0121	0.0096
	(0.0152)	(0.0124)	(0.0121)	(0.0150)	(0.0159)	(0.0161)	(0.0158)	(0.0131)
Population density	0.0002	0.0002*	0.0002	0.0002	0.0002*	0.0002	0.0001	0.0002
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Observations	177	177	177	177	177	177	177	177
R^2	0.1630	0.2132	0.1486	0.1787	0.1573	0.1431	0.1326	0.1468
Adjusted R^2	0.1231	0.1807	0.1133	0.1447	0.1224	0.1076	0.0967	0.1114

The table reports OLS coefficient estimates from a multiple linear regression of absolute mean errors (1992–2018) on the variables shown. Note that the three variable selection methods all select the first nine variables. Column 1 shows the coefficient of freedom status and other variables. Freedom status reflects the Freedom in the World (FiW) index published annually by Freedom House. In this paper, we used the FiW index for 2017. There are three statuses: 1. Not free countries; 2. Partly free countries; 3. Free countries. For other time series variables, I used an average value for 1992–2018. I also employ predictive mean matching to impute missing values (the rate of missing data is shallow). The authors downloaded data from various sources: the WhoGov dataset, Freedom House, Varieties of Democracy dataset, Center of Systemic Peace, World Development Indicator, and Quality of Government. Robust standard errors are in parentheses. $^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

and space. Future research can incorporate cross-sectional differences in the political system, government effectiveness, economic structure, geography, demographic factors, or infrastructure into one model with the new relaxed assumption regarding elasticity. One feasible option is the Bayesian model, which allows for relaxing the original assumption in the lights–GDP association and combines multiple factors in one model. Furthermore, a Bayesian model is more flexible since the research outcome will be a probability density function instead of a single point estimate.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00181-023-02377-y.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Appendix A

See Table 6.

Table 6 Descriptive statistics—data for cross-sectional analysis

Statistic	*N*	Mean	SD	Min	Max
Mean absolute error	172	0.5002	0.3328	0.0367	1.4883
Land area (sq. km)	168	739,286.9000	1,959,270.0000	260.0000	16,389,950.0000
Average precipitation in depth (mm per year) (1992–2010)	164	1,171.9580	798.3112	18.1000	3,240.0000
Average GDP growth (annual %) (1992–2010)	169	3.9886	2.8798	− 1.1985	28.8889
Average GDP per capita growth (annual %) (1992–2010)	169	2.3736	2.6418	− 3.2721	23.7464
Average gross saving (% of GDP) (1992–2010)	153	21.2070	10.9392	− 38.1781	54.4391
Average agriculture, forestry, and fishing, value added (% of GDP) (1992–2010)	168	13.7796	12.5876	0.1957	66.5816
Average industry (including construction), value added (% of GDP) (1992–2010)	167	27.9414	12.4712	5.7734	79.7270
Average services, value added (% of GDP) (1992–2010)	163	51.2838	11.4423	19.3970	77.2517
Average gross capital formation (% of GDP) (1992–2010)	158	23.6171	6.4512	0.0000	48.4352
Average general government final consumption expenditure (% of GDP) (1992–2010)	157	15.8115	5.7695	3.6291	38.0871
Average households and NPISHs final consumption expenditure (% of GDP) (1992–2010)	157	64.9762	17.0667	19.3559	145.5313
Average exports of goods and services (% of GDP) (1992–2010)	162	37.8095	21.1953	7.8515	139.5818

Appendix B

See Table 7.

Table 6 continued

Statistic	*N*	Mean	SD	Min	Max
Average imports of goods and services (% of GDP) (1992–2010)	162	43.2715	20.9769	10.4965	121.3219
Average population growth (annual %) (1992–2010)	170	1.5352	1.3668	− 1.2872	7.8143
Average urban population growth (annual %) (1992–2010)	170	2.2909	1.9141	− 1.3381	8.1490
Average population density (people per sq. km of land area) (1992–2010)	169	119.8267	173.6390	1.5719	1,226.6230
Average population in the largest city (% of urban population) (1992–2010)	138	32.6612	15.8701	3.0510	76.3782
Average agricultural land (% of land area) (1992–2010)	169	38.9109	21.5606	0.5263	84.8366
Average forest area (% of land area) (1992–2010)	169	34.0068	24.7609	0.0000	98.3075
Average number of years the leader in office continuously (1992–2010)	158	7.1378	6.0729	1.0000	31.5000
Average starting a business score (1992–2010)	153	66.0830	19.0286	2.6161	96.8563
Average public sector corruption index (1992–2010)	159	0.4827	0.2949	0.0010	0.9614
Average voice and accountability, estimate (1992–2010)	170	− 0.0096	0.9479	− 1.9397	1.5738
Average control of corruption: estimate (1992–2010)	170	− 0.0310	0.9722	− 1.4856	2.3757
Average State Fragility Index (1995-2010)	152	9.3567	6.4558	0.0000	23.3125
Latitude	167	18.9786	24.6669	− 41.8058	64.9899
Longitude	167	16.1595	62.9711	− 174.8472	171.4777
% Fertile soil	167	39.0006	25.1548	0.0000	100.0000
Percentage desert in 2012	169	3.5625	11.4265	0.0000	77.2795
Percentage tropical climate in 2012	169	41.2543	45.6736	0.0000	100.0000
Ruggedness (Terrain Ruggedness Index, 100 m) in 2012	169	1.3064	1.1914	0.0029	6.7401
Average distance to nearest ice-free coast (1000 km) in 2012	169	0.2968	0.3746	0.00001	2.2062
Percentage within 100 km. of ice-free coast in 2012	169	45.2214	39.8073	0.0000	100.0000
Arable land (% of Agricultural land) (2017)	169	51.5396	29.8284	0.1200	100.0000
Cropland (% of Agricultural land) (2017)	169	39.2047	25.7388	0.1200	98.7700

Table 6 continued

Statistic	*N*	Mean	SD	Min	Max
% European descent	151	33.6827	42.3845	0.0000	100.0000
Ethnic fractionalization in the year 2000	167	0.4401	0.2588	0.0000	0.9302
Language fractionalization in the year 2000	162	0.3924	0.2857	0.0021	0.9227
Religion fractionalization in the year 2000	169	0.4400	0.2375	0.0023	0.8603
Cultural diversity	143	0.3062	0.2072	0.0000	0.7328
Ethnic fractionalization	144	0.4774	0.2626	0.0040	1.0000
Largest minority	136	0.1707	0.1083	0.0100	0.4400
Plurality group	143	0.6456	0.2397	0.1200	0.9980
Perception of Electoral Integrity Index Type	154	3.0779	1.3554	1.0000	5.0000
Legal origin	137	1.9124	0.9194	1.0000	5.0000
Number of years the leader in office continuously in 2010	157	7.2611	7.9358	1.0000	41.0000
Freedom status (2010)	171	2.2339	0.7994	1.0000	3.0000
The region of the country	172	3.5523	1.7109	1	7
Colonial origin	170	1.8235	1.5360	0.0000	4.0000
Natural resource protection indicator	170	69.7548	32.7591	0.6133	100.0000

Table 7 List of countries

Afghanistan	Angola	Albania	United Arab Emirates
Argentina	Armenia	Antigua and Barbuda	Australia
Austria	Azerbaijan	Burundi	Belgium
Benin	Burkina Faso	Bangladesh	Bulgaria
Bahrain	The Bahamas	Bosnia and Herzegovina	Belarus
Belize	Bolivia	Brazil	Barbados
Brunei Darussalam	Bhutan	Botswana	Central African Republic
Canada	Switzerland	Chile	China
Cote d’Ivoire	Cameroon	Congo, Dem. Rep.	Congo, Rep.
Colombia	Comoros	Cabo Verde	Costa Rica
Cyprus	Czech Republic	Germany	Djibouti

Appendix C

See Table 8.

Table 7 continued

Afghanistan	Angola	Albania	United Arab Emirates
Dominica	Denmark	Dominican Republic	Algeria
Ecuador	Egypt, Arab Rep.	Eritrea	Spain
Estonia	Ethiopia	Finland	Fiji
France	Micronesia, Fed. Sts.	Gabon	UK
Georgia	Ghana	Guinea	The Gambia
Guinea-Bissau	Equatorial Guinea	Greece	Grenada
Guatemala	Guyana	Honduras	Croatia
Haiti	Hungary	Indonesia	India
Ireland	Iran, Islamic Rep.	Iraq	Iceland
Israel	Italy	Jamaica	Jordan
Japan	Kazakhstan	Kenya	Kyrgyz Republic
Cambodia	Kiribati	St. Kitts and Nevis	Korea, Rep.
Kuwait	Lao PDR	Lebanon	Liberia
Libya	St. Lucia	Sri Lanka	Lesotho
Lithuania	Luxembourg	Latvia	Morocco
Moldova	Madagascar	Maldives	Mexico
Macedonia, FYR	Mali	Malta	Montenegro
Mongolia	Mozambique	Mauritania	Mauritius
Malawi	Malaysia	Namibia	Niger
Nigeria	Nicaragua	Netherlands	Norway
Nepal	New Zealand	Oman	Pakistan
Panama	Peru	Philippines	Palau
Papua New Guinea	Poland	Portugal	Paraguay
Qatar	Romania	Russian Federation	Rwanda
Saudi Arabia	Sudan	Senegal	Singapore
Saudi Arabia	Sudan	Senegal	Singapore
Solomon Islands	Sierra Leone	El Salvador	Serbia
Sao Tome and Principe	Suriname	Slovak Republic	Slovenia
Sweden	Swaziland	Seychelles	Syrian Arab Republic
Chad	Togo	Thailand	Tajikistan
Turkmenistan	Tonga	Trinidad and Tobago	Tunisia
Turkey	Tanzania	Uganda	Ukraine
Uruguay	USA	Uzbekistan	St. Vincent and the Grenadines
Venezuela, RB	Vietnam	Vanuatu	Samoa
Yemen, Rep.	South Africa	Zambia	

Table 8 Definition and sources of data

Variable	Description	Source
DMSP nighttime light and GDP per capita, PPP (constant 2005 international $)	I use the replicate file of the paper Pinkovskiy and Sala-i Martin (2016). The data cover a panel of 179 countries between 1992 and 2010	Pinkovskiy and Sala-i Martin (2016)
Harmonized global nighttime light dataset (0–63)	Yearly nighttime light data are compiled by Li et al. (2020) and available from the website https://doi.org/10.6084/m9.figshare.9828827.v2. I downloaded GEOTIFF file format and processed raw data in R. The data cover a panel of 208 countries between 1992 and 2018	Li et al. (2020)
GDP per capita, PPP (constant 2017 international $)	GDP per capita based on purchasing power parity (PPP). PPP GDP is gross domestic product converted to international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as the US dollar has in the USA. GDP at purchaser's prices is the sum of gross value added by all resident producers in the country plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in constant 2017 international dollars	WDI
Freedom in the World (FiW) index (2010)	For each country and territory, Freedom in the World analyzes the electoral process, political pluralism and participation, the functioning of the government, freedom of expression and of belief, associational and organizational rights, the rule of law, and personal autonomy and individual rights. Data is available at: https://freedomhouse.org/report/freedom-world	Freedom House
Number of years the leader in office continuously	The number of years the person has been leader of the country in a row. Thus, it starts over if the leader is removed. The count starts at 1, when the leader first appear as leader in the dataset. Therefore, the measure is imprecise for leaders, who came to power before 1966. Available at: https://politicscentre.nuffield.ox.ac.uk/whogov-dataset/download-dataset/	Nyrup and Bramwell (2020)
Public sector corruption index	Question: To what extent do public sector employees grant favors in exchange for bribes, kickbacks, or other material inducements, and how often do they steal, embezzle,or misappropriate public funds or other state resources for personal or family use?. Available at: https://www.v-dem.net/en/data/data/	Varieties of Democracy

Table 8 continued

Variable	Description	Source
Voice and Accountability, Estimate	Voice and Accountability captures perceptions of the extent to which a country's citizens are able to participate in selecting their government, as well as freedom of expression, freedom of association, and a free media. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e., ranging from approximately −2.5 to 2.5	WDI
Starting a business score	The score for starting a business is the simple average of the scores for each of the component indicators: the procedures, time and cost for an entrepreneur to start and formally operate a business, as well as the paid-in minimum capital requirement	WDI
Control of Corruption: Estimate	Control of Corruption captures perceptions of the extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests. Estimate gives the country's score on the aggregate indicator, in units of a standard normal distribution, i.e., ranging from approximately −2.5 to 2.5	WDI
State Fragility Index	A country's fragility is closely associated with its state capacity to manage conflict; make and implement public policy; and deliver essential services and its systemic resilience in maintaining system coherence, cohesion, and quality of life; responding effectively to challenges and crises, and sustaining progressive development. State Fragility = Effectiveness Score + Legitimacy Score (25 points possible). Available at: https://www.systemicpeace.org/inscrdata.html	Center of Systemic Peace
Country Ruggedness and Geographical Data (2012)	The dataset of terrain ruggedness and other geographical characteristics of countries was created by Nathan Nunn and Diego Puga. Available at: https://diegopuga.org/data/rugged/	Nunn and Puga (2012)
Natural resource protection indicator	Natural Resource Protection Indicator assesses whether a country is protecting at least 17% of all of its biomes (e.g., deserts, forests, grasslands, aquatic, and tundra). It is designed to capture the comprehensiveness of a government's commitment to habitat preservation and biodiversity protection. The World Wildlife Fund provides the underlying biome data, and the United Nations Environment Program World Conservation Monitoring Center provides the underlying data on protected areas	Quality of Government (Teorell et al. 2021).

Table 8 continued

Variable	Description	Source
Agriculture, forestry, and fishing, value added (% of GDP)	Agriculture corresponds to ISIC divisions 1–5 and includes forestry, hunting, and fishing, as well as cultivation of crops and livestock production. Value added is the net output of a sector after adding up all outputs and subtracting intermediate inputs. It is calculated without making deductions for depreciation of fabricated assets or depletion and degradation of natural resources. The origin of value added is determined by the International Standard Industrial Classification (ISIC), revision 3. Note: For VAB countries, gross value added at factor cost is used as the denominator	WDI
Industry (including construction), value added (% of GDP)	Industry corresponds to ISIC divisions 10–45 and includes manufacturing (ISIC divisions 15–37). It comprises value added in mining, manufacturing (also reported as a separate subgroup), construction, electricity, water, and gas. Value added is the net output of a sector after adding up all outputs and subtracting intermediate inputs. It is calculated without making deductions for depreciation of fabricated assets or depletion and degradation of natural resources. The origin of value added is determined by the International Standard Industrial Classification (ISIC), revision 3. Note: For VAB countries, gross value added at factor cost is used as the denominator	WDI
Services, value added (% of GDP)	Services correspond to ISIC divisions 50–99 and they include value added in wholesale and retail trade (including hotels and restaurants), transport, and government, financial, professional, and personal services such as education, health care, and real estate services. Also included are imputed bank service charges, import duties, and any statistical discrepancies noted by national compilers as well as discrepancies arising from rescaling. Value added is the net output of a sector after adding up all outputs and subtracting intermediate inputs. It is calculated without making deductions for depreciation of fabricated assets or depletion and degradation of natural resources. The industrial origin of value added is determined by the International Standard Industrial Classification (ISIC), revision 3 or 4	WDI

Table 8 continued

Variable	Description	Source
Manufacturing, value added (% of GDP)	Manufacturing refers to industries belonging to ISIC divisions 15–37. Value added is the net output of a sector after adding up all outputs and subtracting intermediate inputs. It is calculated without making deductions for depreciation of fabricated assets or depletion and degradation of natural resources. The origin of value added is determined by the International Standard Industrial Classification (ISIC), revision 3. Note: For VAB countries, gross value added at factor cost is used as the denominator	WDI
Gross fixed capital formation (% of GDP)	Gross fixed capital formation (formerly gross domestic fixed investment) includes land improvements (fences, ditches, drains, and so on); plant, machinery, and equipment purchases; and the construction of roads, railways, and the like, including schools, offices, hospitals, private residential dwellings, and commercial and industrial buildings. According to the 1993 SNA, net acquisitions of valuables are also considered capital formation	WDI
Gross capital formation (% of GDP)	Gross capital formation (formerly gross domestic investment) consists of outlays on additions to the fixed assets of the economy plus net changes in the level of inventories. Fixed assets include land improvements (fences, ditches, drains, and so on); plant, machinery, and equipment purchases; and the construction of roads, railways, and the like, including schools, offices, hospitals, private residential dwellings, and commercial and industrial buildings. Inventories are stocks of goods held by firms to meet temporary or unexpected fluctuations in production or sales, and "work in progress." According to the 1993 SNA, net acquisitions of valuables are also considered capital formation	WDI
General government final consumption expenditure (% of GDP)	General government final consumption expenditure (formerly general government consumption) includes all government current expenditures for purchases of goods and services (including compensation of employees). It also includes most expenditures on national defense and security, but excludes government military expenditures that are part of government capital formation	WDI

Table 8 continued

Variable	Description	Source
Households and NPISHs final consumption expenditure (% of GDP)	Household final consumption expenditure (formerly private consumption) is the market value of all goods and services, including durable products (such as cars, washing machines, and home computers), purchased by households. It excludes purchases of dwellings but includes imputed rent for owner-occupied dwellings. It also includes payments and fees to governments to obtain permits and licenses. Here, household consumption expenditure includes the expenditures of nonprofit institutions serving households, even when reported separately by the country. This item also includes any statistical discrepancy in the use of resources relative to the supply of resources	WDI
Exports of goods and services (% of GDP)	Exports of goods and services represent the value of all goods and other market services provided to the rest of the world. They include the value of merchandise, freight, insurance, transport, travel, royalties, license fees, and other services, such as communication, construction, financial, information, business, personal, and government services. They exclude compensation of employees and investment income (formerly called factor services) and transfer payments	WDI
Imports of goods and services (% of GDP)	Imports of goods and services represent the value of all goods and other market services received from the rest of the world. They include the value of merchandise, freight, insurance, transport, travel, royalties, license fees, and other services, such as communication, construction, financial, information, business, personal, and government services. They exclude compensation of employees and investment income (formerly called factor services) and transfer payments	WDI
Electric power consumption (kWh per capita)	Electric power consumption measures the production of power plants and combined heat and power plants less transmission, distribution, and transformation losses and own use by heat and power plants	WDI
Access to electricity (% of population)	Access to electricity is the percentage of population with access to electricity. Electrification data are collected from industry, national surveys and international sources	WDI
Oil rents (% of GDP)	Oil rents are the difference between the value of crude oil production at regional prices and total costs of production	WDI

Table 8 continued

Variable	Description	Source
GDP per unit of energy use (constant 2017 PPP $ per kg of oil equivalent)	GDP per unit of energy use is the PPP GDP per kilogram of oil equivalent of energy use. PPP GDP is gross domestic product converted to 2017 constant international dollars using purchasing power parity rates. An international dollar has the same purchasing power over GDP as a US dollar has in the USA	WDI
CO_2 emissions (metric tons per capita)	Carbon dioxide emissions are those stemming from the burning of fossil fuels and the manufacture of cement. They include carbon dioxide produced during consumption of solid, liquid, and gas fuels and gas flaring	WDI
Rail lines (total route—km)	Rail lines are the length of railway route available for train service, irrespective of the number of parallel tracks	WDI
Railways, passengers carried (million passenger—km)	Passengers carried by railway are the number of passengers transported by rail times kilometers traveled	WDI
Automated teller machines (ATMs) (per 100,000 adults)	Automated teller machines are computerized telecommunications devices that provide clients of a financial institution with access to financial transactions in a public place	WDI
Mobile cellular subscriptions (per 100 people)	Mobile cellular telephone subscriptions are subscriptions to a public mobile telephone service that provide access to the PSTN using cellular technology. The indicator includes (and is split into) the number of postpaid subscriptions, and the number of active prepaid accounts (i.e., that have been used during the last 3 months). The indicator applies to all mobile cellular subscriptions that offer voice communications. It excludes subscriptions via data cards or USB modems, subscriptions to public mobile data services, private trunked mobile radio, telepoint, radio paging, and telemetry services	WDI
Fixed telephone subscriptions (per 100 people)	Fixed telephone subscriptions refers to the sum of active number of analog fixed telephone lines, voice over IP (VoIP) subscriptions, fixed wireless local loop (WLL) subscriptions, ISDN voice-channel equivalents and fixed public payphones	WDI
Container port traffic (TEU: 20 foot equivalent units)	Port container traffic measures the flow of containers from land to sea transport modes, and vice versa, in twenty-foot equivalent units (TEUs), a standard-size container. Data refer to coastal shipping as well as international journeys. Transshipment traffic is counted as two lifts at the intermediate port (once to off-load and again as an outbound lift) and includes empty units	WDI

Table 8 continued

Variable	Description	Source
Air transport, registered carrier departures worldwide	Registered carrier departures worldwide are domestic takeoffs and takeoffs abroad of air carriers registered in the country	WDI
Population density (people per sq. km of land area)	Population density is midyear population divided by land area in square kilometers. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship–except for refugees not permanently settled in the country of asylum, who are generally considered part of the population of their country of origin. Land area is a country's total area, excluding area under inland water bodies, national claims to continental shelf, and exclusive economic zones. In most cases the definition of inland water bodies includes major rivers and lakes	WDI
Population in the largest city (% of urban population)	Population in largest city is the percentage of a country's urban population living in that country's largest metropolitan area	WDI
Agricultural land (% of land area)	Agricultural land refers to the share of land area that is arable, under permanent crops, and under permanent pastures. Arable land includes land defined by the FAO as land under temporary crops (double-cropped areas are counted once), temporary meadows for mowing or for pasture, land under market or kitchen gardens, and land temporarily fallow. Land abandoned as a result of shifting cultivation is excluded. Land under permanent crops is land cultivated with crops that occupy the land for long periods and need not be replanted after each harvest, such as cocoa, coffee, and rubber. This category includes land under flowering shrubs, fruit trees, nut trees, and vines, but excludes land under trees grown for wood or timber. Permanent pasture is land used for five or more years for forage, including natural and cultivated crops	WDI
Forest area (% of land area)	Forest area is land under natural or planted stands of trees of at least 5 meters in situ, whether productive or not, and excludes tree stands in agricultural production systems (for example, in fruit plantations and agroforestry systems) and trees in urban parks and gardens	WDI
Permanent cropland (% of land area)	Permanent cropland is land cultivated with crops that occupy the land for long periods and need not be replanted after each harvest, such as cocoa, coffee, and rubber. This category includes land under flowering shrubs, fruit trees, nut trees, and vines, but excludes land under trees grown for wood or timber	WDI

Table 8 continued

Variable	Description	Source
Foreign direct investment, net inflows (% of GDP)	Foreign direct investment are the net inflows of investment to acquire a lasting management interest (10 percent or more of voting stock) in an enterprise operating in an economy other than that of the investor. It is the sum of equity capital, reinvestment of earnings, other long-term capital, and short-term capital as shown in the balance of payments. This series shows net inflows (new investment inflows less disinvestment) in the reporting economy from foreign investors, and is divided by GDP	WDI
Net ODA received (% of central government expense)	Net official development assistance (ODA) consists of disbursements of loans made on concessional terms (net of repayments of principal) and grants by official agencies of the members of the Development Assistance Committee (DAC), by multilateral institutions, and by non-DAC countries to promote economic development and welfare in countries and territories in the DAC list of ODA recipients. It includes loans with a grant element of at least 25 percent (calculated at a rate of discount of 10 percent)	WDI
Net ODA received (% of gross capital formation)	Net official development assistance (ODA) consists of disbursements of loans made on concessional terms (net of repayments of principal) and grants by official agencies of the members of the Development Assistance Committee (DAC), by multilateral institutions, and by non-DAC countries to promote economic development and welfare in countries and territories in the DAC list of ODA recipients. It includes loans with a grant element of at least 25 percent (calculated at a rate of discount of 10 percent)	WDI
Personal remittances, received (% of GDP)	Personal remittances comprise personal transfers and compensation of employees. Personal transfers consist of all current transfers in cash or in kind made or received by resident households to or from nonresident households. Personal transfers thus include all current transfers between resident and nonresident individuals. Compensation of employees refers to the income of border, seasonal, and other short-term workers who are employed in an economy where they are not resident and of residents employed by nonresident entities. Data are the sum of two items defined in the sixth edition of the IMF's Balance of Payments Manual: personal transfers and compensation of employees	WDI

Table 8 continued

Variable	Description	Source
Statistical Capacity score (Overall average)	The Statistical Capacity Indicator is a composite score assessing the capacity of a country's statistical system. It is based on a diagnostic framework assessing the following areas: methodology; data sources; and periodicity and timeliness. Countries are scored against 25 criteria in these areas, using publicly available information and/or country input. The overall Statistical Capacity score is then calculated as a simple average of all three area scores on a scale of 0–100	WDI
GDP growth (annual %)	Annual percentage growth rate of GDP at market prices based on constant local currency. Aggregates are based on constant 2010 US dollars. GDP is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources	WDI
GDP per capita growth (annual %)	Annual percentage growth rate of GDP per capita based on constant local currency. Aggregates are based on constant 2010 US dollars. GDP per capita is gross domestic product divided by midyear population. GDP at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources	WDI
Gross savings (% of GDP)	Gross savings are calculated as gross national income less total consumption, plus net transfers	WDI
Latitude	Expressed in decimal degress, for the geographical centroid of the country	(Nunn and Puga 2012)
Longitude	Expressed in decimal degress, for the geographical centroid of the country	(Nunn and Puga 2012)
% European descent	The variable, calculated from version 1.1 of the migration matrix of Putterman and Weil (2010), estimates the percentage of the year 2000 population in every country that is descended from people who resided in Europe in 1500	(Nunn and Puga 2012)

Table 8 continued

Variable	Description	Source
Average distance to nearest ice-free coast (1000 km)	To calculate the average distance to the closest ice-free coast in each country, we first compute the distance to the nearest ice-free coast for every point in the country in equi-rectangular projection with standard parallels at 30 degrees, on the basis of sea and sea ice area features contained in the fifth edition of the Digital Chart of the World (US National Imagery and Mapping Agency, 2000) and the country boundaries described above. We then average this distance across all land in each country not covered by inland water features. Units are thousands of kilometers	(Nunn and Puga 2012)
Ethnic Fractionalization in the year 2000	The definition of ethnicity involves a combination of racial and linguistic characteristics. The result is a higher degree of fractionalization than the commonly used ELF-index (see el_elf60) in for example Latin America, where people of many races speak the same language	Quality of Government (Teorell et al. 2021)
Language Fractionalization in the year 2000	Linguistic Fractionalization in the year 2000. Reflects probability that two randomly selected people from a given country will not belong to the same linguistic group. The higher the number, the more fractionalized society	Quality of Government (Teorell et al. 2021)
Religion Fractionalization in the year 2000	Religious Fractionalization in the year 2000. Reflects probability that two randomly selected people from a given country will not belong to the same religious group. The higher the number, the more fractionalized society	Quality of Government (Teorell et al. 2021)
Cultural Diversity	This measure modifies fractionalization (fe_etfra) so as to take some account of cultural distances between groups, measured as the structural distance between languages spoken by different groups in a country. If the groups in a country speak structurally unrelated languages, their cultural diversity index will be the same as their level of ethnic fractionalization (fe_etfra). The more similar are the languages spoken by different ethnic groups; however, the more will this measure be reduced below the level of ethnic fractionalization for that country. The values are assumed to be constant for all years	Quality of Government (Teorell et al. 2021)
Ethnic Fractionalization	Restricting attention to groups that had at least 1 percent of country population in the 1990s, Fearon identifies 822 ethnic and "ethnoreligious" groups in 160 countries. This variable reflects the probability that two randomly selected people from a given country will belong to different such groups. The variable thus ranges from 0 (perfectly homogeneous) to 1 (highly fragmented). The values are assumed to be constant for all years	Quality of Government (Teorell et al. 2021)

Table 8 continued

Variable	Description	Source
Largest Minority	Based on the same set of groups, this variable reflects the population share of the second largest group (largest minority). The values are assumed to be constant for all years	Quality of Government (Teorell et al. 2021)
Plurality Group	Based on the same set of groups, this variable reflects the population share of the largest group (plurality group) in the country. The values are assumed to be constant for all years	Quality of Government (Teorell et al. 2021)
Legal origin	Identifies the legal origin of the Company Law or Commercial code of each country. There are five possible origins: (1) English Common Law, (2) French Commercial Code, (3) Socialist/Communist Laws, (4) German Commercial Code, (5) Scandinavian Commercial Code	Quality of Government (Teorell et al. 2021)
Colonial Origin	This is a tenfold classification of the former colonial ruler of the country. Following Bernard et al. (2004), we have excluded the British settler colonies (the USA, Canada, Australia, Israel, and New Zealand), and exclusively focused on "Western overseas" colonialism. This implies that only Western colonizers (e.g., excluding Japanese colonialism), and only countries located in the non-Western hemisphere "overseas," e.g., excluding Ireland & Malta), have been coded. Each country that has been colonized since 1700 is coded. In cases of several colonial powers, the last one is counted, if it lasted for 10 years or longer	Quality of Government (Teorell et al. 2021)
Urban population (% of total population)	Urban population refers to people living in urban areas as defined by national statistical offices. The data are collected and smoothed by United Nations Population Division	WDI
Rural population (% of total population)	Rural population refers to people living in rural areas as defined by national statistical offices. It is calculated as the difference between total population and urban population	WDI
Other variables	Other variables are described in detail in the codebook of the Quality of Government Standard Dataset (Teorell et al. 2021)	

References

Asher S, Lunt T, Matsuura R, Novosad P (2021) Development research at high geographic resolution: an analysis of night-lights, firms, and poverty in India using the shrug open data platform

Breheny P (2016) Adaptive lasso, MCP, and SCAD

Breheny P, Huang J (2011) Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. Ann Appl Stat 5(1):232

Bühlmann P, Van De Geer S (2011) Statistics for high-dimensional data: methods, theory and applications. Springer, Berlin

Chen X, Nordhaus WD (2011) Using luminosity data as a proxy for economic statistics. Proc Natl Acad Sci 108(21):8589–8594

Chen X, Nordhaus WD (2019) Viirs nighttime lights in the estimation of cross-sectional and time-series GDP. Remote Sens 11(9):1057

Doll CH, Muller J-P, Elvidge CD (2000) Night-time imagery as a tool for global mapping of socioeconomic parameters and greenhouse gas emissions. AMBIO J Hum Environ 29(3):157–162

Feige EL, Urban I (2008) Measuring underground (unobserved, non-observed, unrecorded) economies in transition countries: can we trust GDP? J Comp Econ 36(2):287–306

Ghosh T, Powell RL, Elvidge CD, Baugh KE, Sutton PC, Anderson S (2010) Shedding light on the global distribution of economic activity. Open Geogr J 3(1) 17–28

Gibson J, Olivia S, Boe-Gibson G, Li C (2021) Which night lights data should we use in economics, and where? J Dev Econ 149:102602

Henderson JV, Storeygard A, Weil DN (2012) Measuring economic growth from outer space. Am Econ Rev 102(2):994–1028

Hodler R, Raschky P (2014) Regional favoritism. Q J Econ 129(2):995–1033

Hu Y, Yao J (2021) Illuminating economic growth. J Econom 228:359–378

Ishwaran H, Rao JS (2005) Spike and slab variable selection: frequentist and bayesian strategies. Ann Stat 33(2):730–773

Keola S, Andersson M, Hall O (2015) Monitoring economic development from space: using nighttime light and land cover data to measure economic growth. World Dev 66:322–334

Li X, Zhou Y, Zhao M, Zhao X (2020) A harmonized global nighttime light dataset 1992–2018. Sci. Data 7(1):1–9

Marshall MG, Gurr TR, Jaggers K (2011) Center for systemic peace. Polity IV Project

Martinez LR (2022) How much should we trust the dictator's GDP growth estimates? J Polit Econ 130(10):2731–2769

Nunn N, Puga D (2012) Ruggedness: The blessing of bad geography in Africa. Rev Econ Stat 94(1):20–36

Nyrup J, Bramwell S (2020) Who governs? A new global dataset on members of cabinets. Am Polit Sci Rev 114(4):1366–1374

Pinkovskiy M, Sala-i Martin X (2016) Lights, camera... income! illuminating the national accounts-household surveys debate. Q J Econ 131(2):579–631

Teorell J, Sundström A, Holmberg S, Rothstein B, Alvarado Pachon N, Dalli CM (2021) The quality of government standard dataset, version Jan21. University of Gothenburg: The Quality of Government Institute, Gothenburg

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol) 58(1):267–288

Varian HR (2014) Big data: new tricks for econometrics. J Econ Perspect 28(2):3–28

Wu J, Wang Z, Li W, Peng J (2013) Exploring factors affecting the relationship between light consumption and GDP based on DMSP/OLS nighttime satellite imagery. Remote Sens Environ 134:111–119

Zhang C-H (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Stat 38(2):894–942

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.