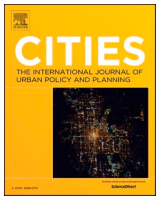
Cities 166 (2025) 106184

Contents lists available at ScienceDirect

Cities

journal homepage: www.elsevier.com/locate/cities



The determinants of national city size distributions: A BMA approach☆

Miguel Puente-Ajovín [a], Marcos Sanso-Navarro [a], María Vera-Cabello [b,*]

[a] Departamento de Análisis Económico & IEDIS, Universidad de Zaragoza, Spain
[b] Centro Universitario de la Defensa de Zaragoza, Spain

ARTICLE INFO

JEL classifications:
O10
O18
O57
R12

Keywords:
City size distribution
Gridded population
Nighttime lights
Bayesian model averaging

ABSTRACT

This paper investigates the factors that influence the national distribution of city sizes, using data from both gridded population estimates and nighttime lights. Leveraging a global definition of human settlements, it also accounts for spatial units smaller than traditional urban centers. To address model uncertainty in identifying robust determinants of cross-country variation in urban concentration, the analysis employs Bayesian model averaging techniques. Across different samples, openness to trade, the Polity score, and natural resource dependence consistently show posterior inclusion probabilities of approximately 50 % or higher. The findings highlight the potential roles of international trade and institutional quality in fostering more balanced national urban systems.

1. Introduction

The percentage of urban population worldwide surpassed that of rural population for the first time in 2007 (Wimberley, Morris, & Fulkerson, 2007). After that moment, the share of global urban population has steadily increased, currently standing at 56 %, according to the World Bank[1] (An, Choi, Lee, Kim, & Lee, 2024). Furthermore, this institution expects that seven out of ten people will reside in cities by 2050. With over 80 % of the global gross domestic product (GDP) generated within cities, the study of their developmental trajectories – acknowledging that the growth rate of urban centers is not uniform (Duranton & Puga, 2014) – and the resulting distribution of city sizes becomes increasingly pertinent. Indeed, this distribution is one of the most extensively researched topics among urban economists due to its significant theoretical and policy implications.

Building on the groundbreaking work of Gabaix (1999) and Eeckhout (2004), subsequent research on the distribution of city sizes has predominantly focused on its conformity to the rank-size rule – commonly referred to as Zipf's law – and has primarily examined the urban structure within single countries. The relatively smaller number of cross-country studies can be attributed to variations in the definition of what constitutes a city across different national data sources, which limits comparability. This highlights the narrow scope of existing research on the factors influencing the city size distribution, even if it provides insights into the spatial organization of economic activity, population dynamics, and the efficiency of urban systems (Krugman, 1991). The way cities are distributed in size – whether dominated by a single large metropolis or more evenly spread across several mid-sized cities – can have profound implications for economic development, regional inequality, and infrastructure provision (Henderson, 2003). In particular, more balanced urban structures are often associated with broader access to public goods, reduced regional disparities, and more resilient national economies. Therefore, the knowledge of the drivers of city size distributions is useful for the design of spatial development strategies supporting inclusive and sustainable urbanization.

An early study of the determinants of the distribution of city sizes at the country level was conducted by Rosen and Resnick (1980), who demonstrated that estimated coefficients from national rank-size regressions – a widely used measure of urban concentration – are directly (inversely) related to GDP per capita, total population, and railroad

☆ The authors have benefited from the valuable comments of the Editor (Pengjun Zhao), three anonymous reviewers, and participants at the XLVIII International Conference on Regional Science (Cuenca), and the 71st North American Meetings of the Regional Science Association International (New Orleans). This work has been supported by Gobierno de Aragón (S39_23R ADETRE Research Group) and Ministerio de Ciencia e Innovación/Agencia Estatal de Investigación (Grant PID2020-112773GB-I00).

* Corresponding author at: Academia General Militar, Carretera de Huesca s/n, 50090 Zaragoza, Spain.
E-mail address: mvera@unizar.es (M. Vera-Cabello).

[1] https://www.worldbank.org/en/topic/urbandevelopment/overview.

https://doi.org/10.1016/j.cities.2025.106184
Received 17 July 2024; Received in revised form 15 April 2025; Accepted 15 June 2025
Available online 27 June 2025

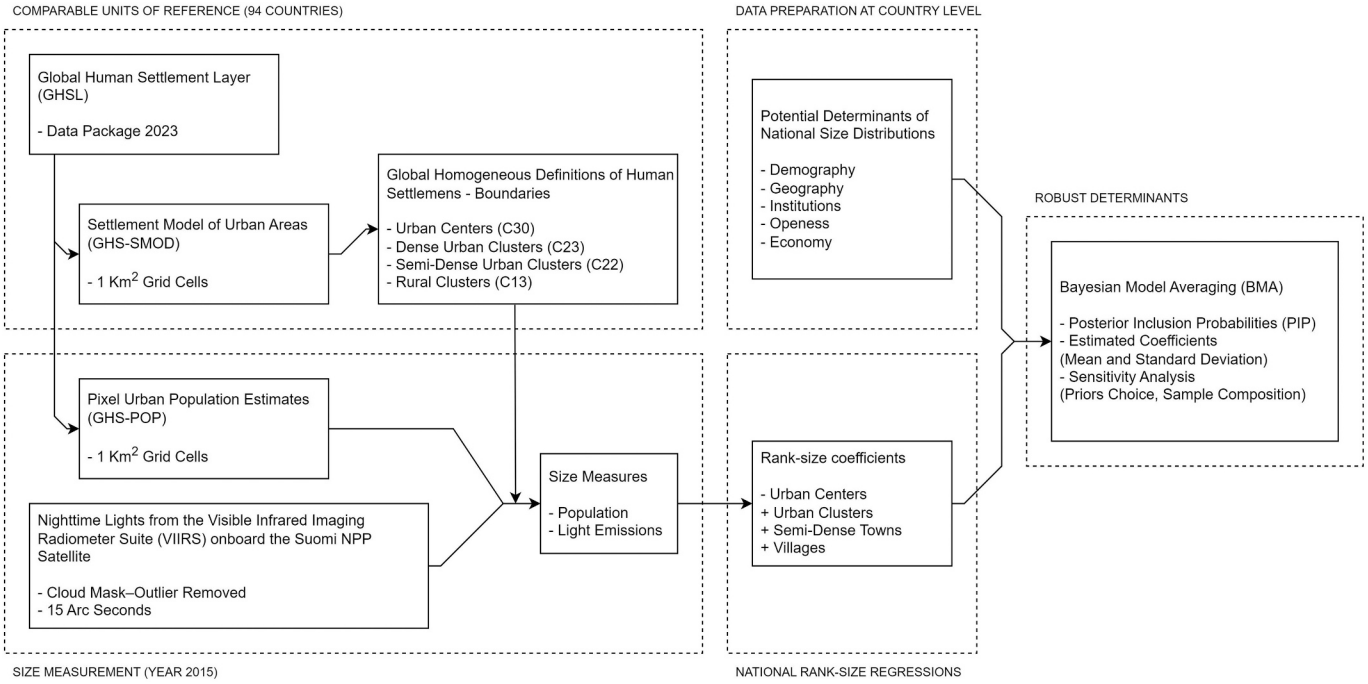


Fig. 1. Conceptual framework: A BMA approach to the determinants of national size distributions.

density (land area). Using an expanded sample, Soo (2005) concluded that political factors play a more significant role than economic variables in shaping the size distribution of cities. More recently, Modica (2017) examines the impact of European Union integration on the distribution of city sizes. His findings indicate that, while the Schengen Agreement led to greater urban population concentration, the introduction of the common European currency contributed to increased dispersion. He also identifies a positive (negative) correlation between urban concentration and total population (land area). Similarly, Sun et al. (2021) investigate the relationship between socioeconomic factors and city size distribution, concluding that polarization should not be a primary concern. They highlight factors such as urbanization, declining state fragility, and advancements in internet technology, which disproportionately benefit smaller cities.

To address the challenge of comparability, Wang, Sun, and Zhang (2021) identify physical cities using a harmonized global definition based on LandScan population data and the European Space Agency land cover map. Their results indicate the existence of an inverted U-shaped relationship between internet penetration rates and the equality of national city size distributions; see Wang, Zhou, and Sun (2022) for a related study with a broader sample. Using the same city definition, Wang, Wei, and Sun (2022) examine the impact of the "new economy" on national city size distributions. Their findings suggest that, while innovation generally fosters a more egalitarian distribution, the influence of human capital varies across country types, and the effects of globalization and information and communication technologies fluctuate over time.

Düben and Krause (2021) employ geospatial data from a globally consistent city identification framework based on the Global Human Settlement Layer (GHSL) and use nighttime lights (NTL) as a proxy for economic activity (Chen & Nordhaus, 2011; Henderson, Storeygard, & Weil, 2012). In addition to comparing the distribution of urban population with that of light emissions as a means of analyzing the magnitude of agglomeration economies, the authors also investigate the determinants of cross-country heterogeneity in city size distributions. To this end, they apply a model selection approach based on a simplified algorithm that evaluates all combinations of up to seven regressors, ultimately concluding that historical factors play a significant role – interpreted as supporting evidence for the "time of development" hypothesis.

As an alternative approach to addressing model uncertainty in the analysis of the determinants of national city size distributions, we propose the implementation of Bayesian model averaging (BMA) techniques (Raftery, Madigan, & Hoeting, 1997). BMA provides a flexible framework that does not require fixing the model size or selecting a specific subset of regressors ex ante, thereby allowing for the evaluation of a wide range of potential explanatory variables. This flexibility is particularly valuable in a cross-sectional setting, where the number of predictors is large relative to the number of observations. By estimating all candidate models and averaging their results using posterior probabilities as weights, BMA accounts for uncertainty both within and across models. This integrated approach to model selection, estimation, and inference will enable us to draw more robust conclusions about the factors associated with urban concentration.

Although the literature on the determinants of city size distributions has primarily focused on large cities or metropolitan areas, there is growing recognition of the importance of incorporating smaller urban settlements into the analysis. As noted by Christiaensen and Todo (2014), these spatial units often play critical roles in national and regional economies, serving as hubs for local trade, service provision, and rural-urban linkages. Ignoring such settlements can therefore lead to an incomplete understanding of urban systems (Duranton, 2015; Eeckhout, 2004). To enable a more comprehensive assessment of the determinants of national urban structure, we adopt a globally consistent definition of human settlements and include smaller spatial units in our analysis to identify variables that are strongly associated with the degree of urban concentration across countries. The conceptual framework guiding the study is presented in Fig. 1.

The remainder of the paper is structured as follows. Section 2 describes the spatial units considered in the analysis and the georeferenced data used to calculate their sizes. Section 3 outlines the estimation of the concentration measure under study. Section 4 discusses its potential national-level determinants, while Section 5 details the BMA techniques employed to assess their relevance. Section 6 presents the main findings, evaluates their sensitivity to prior specifications and the composition of national samples, and compares the results with those obtained using

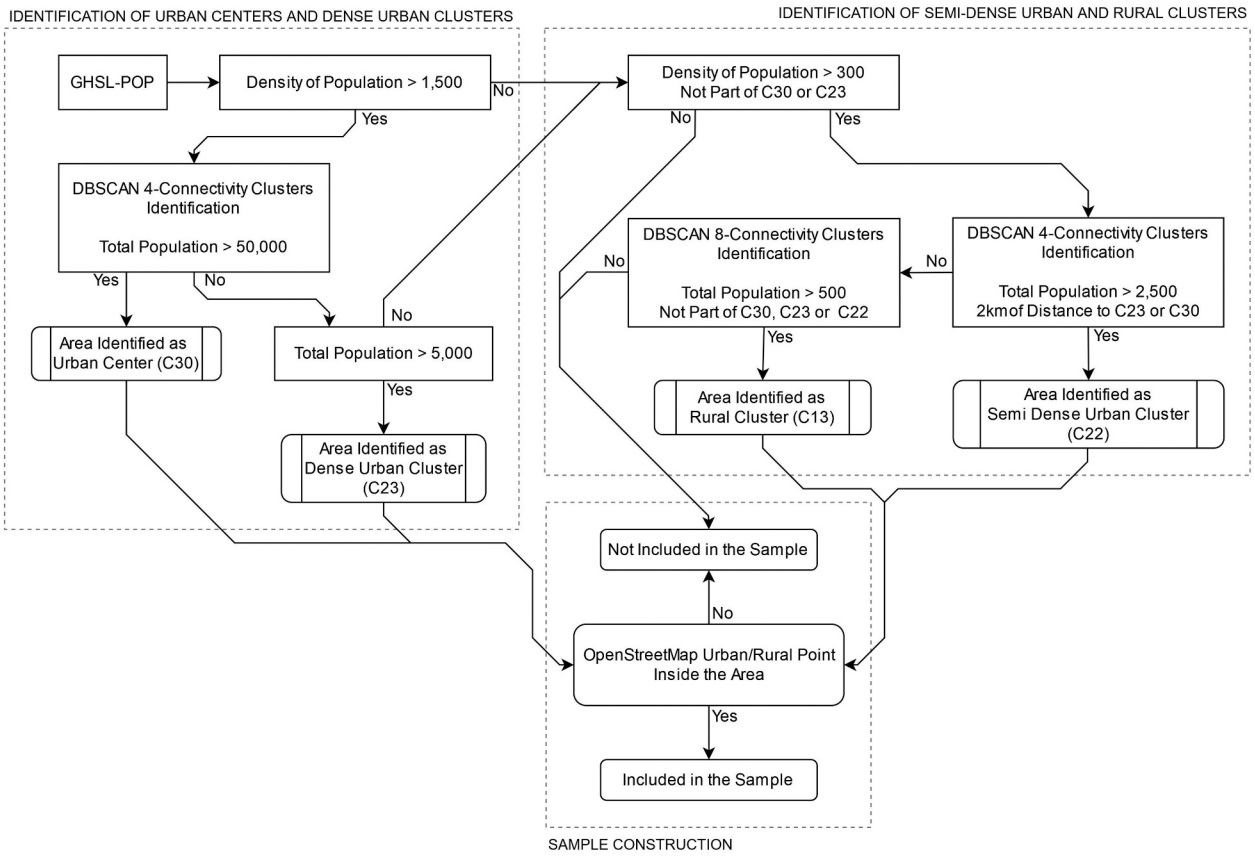


Fig. 2. GHSL-POP population clusters identification and sample construction.

traditional linear regression models. Finally, Section 7 concludes. An appendix provides additional information and results.

2. City identification and size measurement

The study of the determinants of the national size distribution of cities requires a consistent definition of these entities across countries. To ensure comparability and as an initial step, we identified cities using data from the GHSL (Data Package 2023), an initiative of the Joint Research Center of the European Commission (Florczyk et al., 2019; Florczyk et al., 2019). The Settlement Model of Urban Areas in this database (GHS-SMOD) combines information on built-up areas (derived from Landsat imagery) with gridded population data (Gridded Population of the World, Version 4), and divides the globe into one-square-kilometer grid cells, classifying them as part of a rural area or an urban center[2] and/or an urban cluster.

The sample analyzed in this study was constructed from the different clusters identified in the GHSL population grid (GHS-POP) for 2015 using the DBSCAN algorithm (Ester, Kriegel, Sander, & Xiaowei, 1996), as well as population density and total population data, in alignment with recommendations for classifying human settlements as urban in both developed and developing countries (World Bank, 2008). The first category considered, adopted as the reference unit, is 'urban centers (C30)', defined as areas with a population density of at least $1,500$ people per square kilometer (km^2) and a total population of $50,000$ inhabitants or more. Additionally, two other urban subcategories are included: 'dense urban clusters (C23)', which have a minimum density of $1,500$ inhabitants per km^2 and a total population exceeding $50,000$, and 'semi-dense urban clusters (C22)', characterized by a minimum density of 300 inhabitants per km^2 and a total population of over $2,500$. For rural areas, the analysis also considers those classified as 'rural clusters (C13)', which do not fall under any urban category but have a population density exceeding 300 people per km^2 and at least 500 inhabitants.[3]

It is worth noting that while urban areas are included in 'urban centers' and 'dense urban clusters', 'semi-dense urban clusters' and 'rural clusters' may contain industrial zones, airports, ports, or other built-up areas that, although designated as urban or rural entities, do not fully conform to their respective characteristics. To address these anomalies, we have incorporated geographical coordinate data for cities, towns, villages, and hamlets worldwide, extracted from OpenStreetMap. We postulated that any area not encompassing one of these

[2] The concept of urban centers established by the GHSL aims to capture the functionality and density of urban areas in a globally consistent manner, while the definition of a city remains shaped by local administrative frameworks. In terms of population, the GHSL focuses on dense urban clusters, which may exclude areas that are legally or administratively classified as cities but do not meet the minimum density threshold. Conversely, areas not classified as cities by legal standards may still be regarded as urban centers by the GHSL if they meet its criteria for density and size. Regarding boundaries, the GHSL defines those for urban centers more flexibly, based on the physical distribution of the population, whereas city boundaries are determined by fixed administrative limits. This latter distinction makes identifying cities based on administrative definitions relatively stable, regardless of the emergence of genuinely new urban centers (Wang & Sun, 2024). Consequently, an urban center may encompass areas that are not officially cities or may extend beyond the boundaries of a given city.

[3] Other subcategories – 'suburban', 'peri-urban', 'low-density rural', and 'very low-density rural' grid cells – were excluded because the GHS-SMOD does not classify them as entities. These categories lack a minimum population size threshold and are characterized by very low population densities.

Table 1
Descriptive statistics: Size of GHSL human settlements, year 2015.

	Population			
	Urban centers	Urban clusters	Semi-dense towns	Villages
Observations	10,686	55,991	22,277	190,532
Mean	295,575.65	14,488.78	8567.05	1623.04
Median	103,515.97	10,375.77	6808.27	1223.14
Minimum	20,158.45	1166.98	1123.44	11.83
Maximum	37,856,927.80	100,868.17	347,085.38	9423.50

	Light emissions			
	Urban centers	Urban clusters	Semi-dense towns	Villages
Observations	10,357	49,480	18,103	130,044
Mean	6660.77	272.53	241.42	45.86
Median	979.20	53.65	60.23	12.94
Minimum	0.09	0.01	0.06	0.01
Maximum	1,132,904.24	174,562.75	29,939.47	14,142.55

Note: The number of countries included in the sample is 94. Population is measured in number of persons within the corresponding spatial extents. Light emissions are expressed as aggregate nano Watts per square centimeter per steredian.

database. For urban size measured in demographic terms, we used pixel-level population estimates from GHS-POP[4] for the year 2015. To calculate a measure of urban economic activity, we build on the work of Puente-Ajovin, Sanso-Navarro, and Vera-Cabello (2022) and Puente-Ajovin, Sanso-Navarro, and Vera-Cabello (2024), employing the more precise NTL data from the Visible Infrared Imaging Radiometer Suite (VIIRS) aboard the Suomi NPP satellite.[5] Specifically, we extracted the 'vcm-orm-ntl' annual composites[6] for 2015 from the website of the Earth Observation Group of the National Oceanic and Atmospheric Administration (US Department of Commerce). Although the VIIRS data pixels are smaller than those of GHSL, this discrepancy is not problematic, as we aggregated the light emissions based on the larger GHSL pixels. Table 1 presents descriptive statistics for the sizes of human settlements across the 94 countries in our sample, calculated according to the alternative definitions described above.[7]

3. Rank-size regressions at the country level

The rank-size rule suggests that the city size distribution can be approximated by a Pareto function with a power law exponent equal to one. Therefore, cross-sectional empirical analyses of Zipf's law typically rely on a log-log linear regression between the rank of a city and its size. To mitigate the small-sample biases of ordinary least squares (OLS) regressions, Gabaix and Ibragimov (2011) propose estimating the following model:

$$log(Rank_i - 0.5) = \alpha - \beta \cdot log(Size_i) + \varepsilon_i, \quad i = 1, \ldots, n; \tag{1}$$

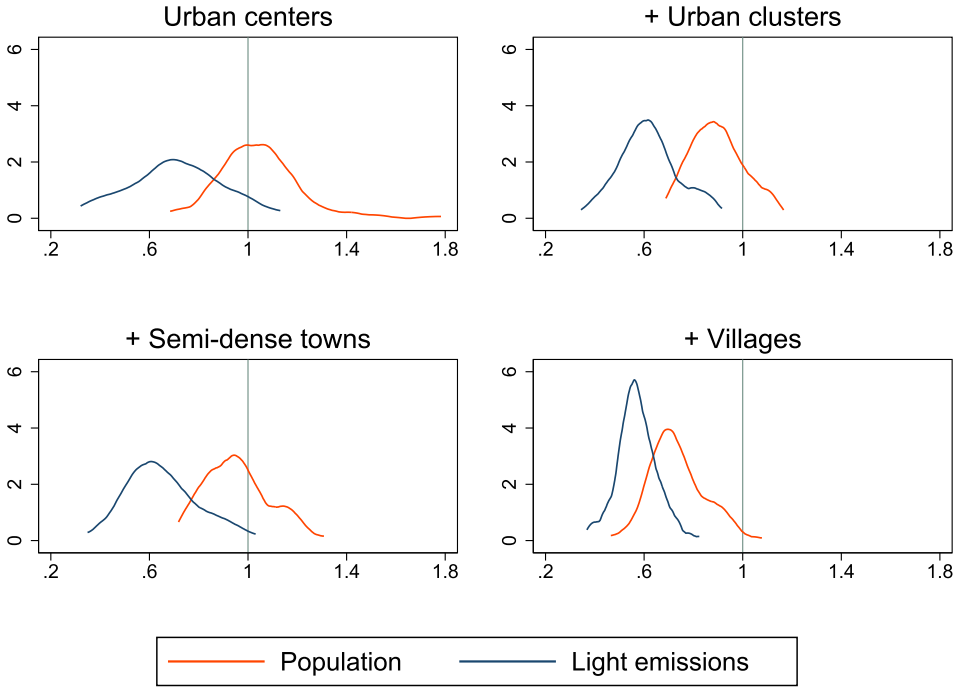


Fig. 3. Kernel densities of estimated coefficients from national rank-size OLS regressions: Alternative sample compositions.

points should not be classified as urban or rural and, therefore, excluded it from our sample. To account for potential discrepancies in the precise localization of these points, we retained polygons that, while not containing a point inside, had one in an adjacent pixel. This approach ensures a more robust and accurate representation of urban and rural areas. The technicalities in the definition of the categories considered in our sample, and the process followed to select the units included in it are illustrated in Fig. 2.

Although the vast majority of studies on city size distribution rely on population data, we additionally use NTL satellite imagery as a proxy for urban economic activity. Consequently, city size will be measured by the sum of the population and aggregate light emissions within the spatial extent of GHSL units, as defined by the shapefile provided by this

[4] See Puente-Ajovin, Sanso-Navarro, and Vera-Cabello (2024), as well as the references therein, for a discussion of alternative gridded population databases.
[5] See also Appiah-Kubi and Gyambibi (2025) for a recent related contribution focusing on African countries.
[6] VIIRS Cloud Mask–Outlier Removed–Nighttime Lights.
[7] Refer to Table A1 in the appendix for additional details on the composition of the analyzed sample.

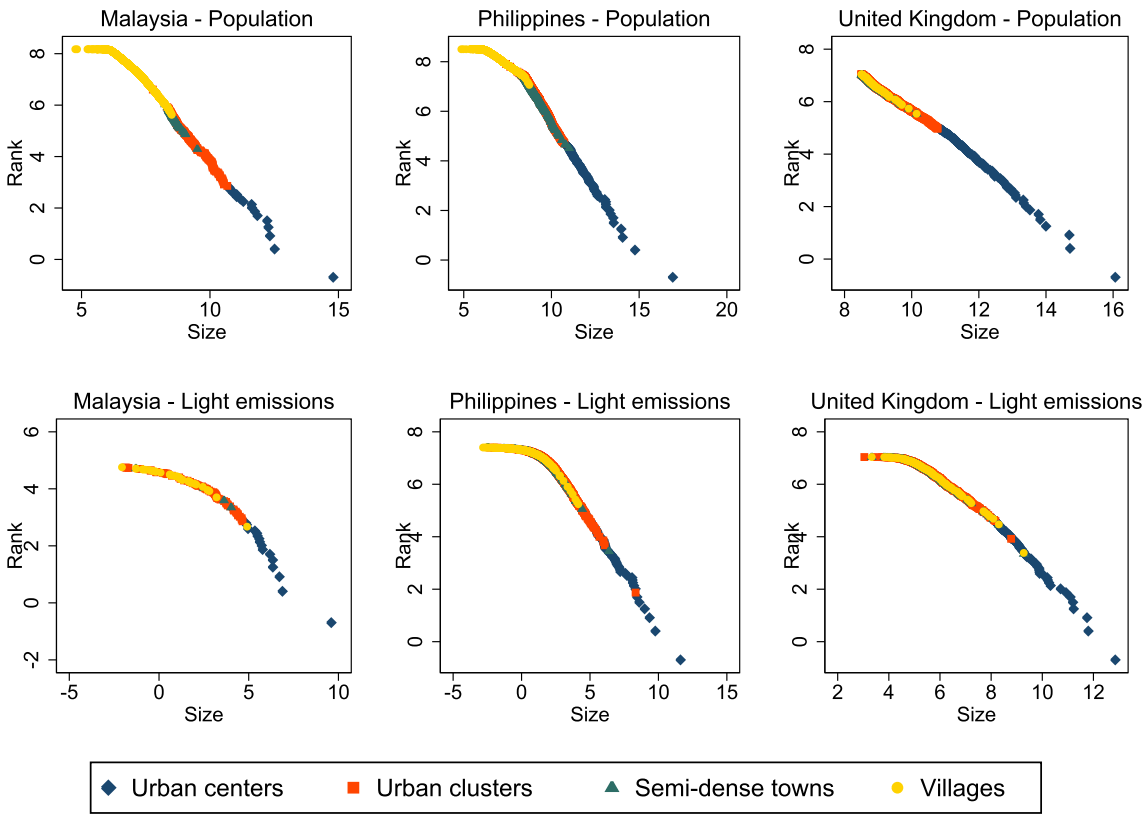


Fig. 4. Rank-size (in logarithms) plots for selected countries and alternative sample compositions.

where i denotes a city indicator, and n represents the sample size. Zipf's law corresponds to $\beta = 1$. In our context, a coefficient less (greater) than one indicates that the population and/or light emissions are more unequally (equally) distributed across the national urban system than predicted by the rank-size rule.

Fig. 3 presents kernel density estimates for the slope parameter in expression (1) – commonly referred to as Pareto coefficient – at the country level,[8] where city size is measured in demographic terms (orange) and urban economic activity is proxied by aggregate light emissions (blue). In national samples limited to urban centers (baseline), rank-size coefficients tend to cluster around one when city sizes are measured by population. In contrast, slope parameters estimated using NTL typically yield values below unity. These findings align with prior literature, which suggests that urban light emissions are more unevenly distributed than population at the country level (Düben & Krause, 2021; Puente-Ajovin et al., 2024), highlighting the global significance of urban agglomeration economies. Examining the data in greater detail,[9] the highest estimated coefficients for city sizes measured demographically are observed in Ethiopia (1.78) and the Democratic People's Republic of Korea (1.52), while the lowest are found in Greece (0.68) and Austria (0.71). For national rank-size regressions based on aggregate urban NTL, the highest coefficients are recorded in France (1.13) and Romania (1.09), whereas the lowest are in Sudan (0.32) and Azerbaijan (0.33).

The remaining graphs in Fig. 3 illustrate that including human settlements with less urban characteristics in the sample generally reduces the estimated slope parameters. This effect is particularly pronounced when size is measured using gridded population data and when villages are considered. In fact, this expanded sample composition results in the greatest similarity between the distributions of population and NTL. This outcome shows that accounting for smaller human settlements uncovers the uneven distribution of population at the country level while simultaneously diminishing evidence for agglomeration economies. While we will further analyze how conclusions about the robust determinants of national size distributions change as smaller settlements are included, it is important to emphasize that the validity of Pareto coefficients relies on the assumption that sizes follow this distribution. The prevailing consensus in the literature is that city size distributions are better modeled as a mixture of a lognormal body with an upper Pareto tail; see Puente-Ajovin, Ramos, and Sanz-Gracia (2020) and references therein. As reflected in Fig. 3, and as noted by Eeckhout (2004), estimated coefficients from rank-size regressions systematically decline with increasing sample size when the underlying distribution is lognormal rather than Pareto; see Wang and Sun (2024) for a recent related contribution.

Fig. 4 illustrates that adding increasingly smaller cities to a national sample results in growing inequality in the size distribution. This is demonstrated for three selected countries according to their estimated rank-size coefficients for the population of urban centers: Malaysia (Q1, 0.92), the Philippines (Median, 1.03), and the United Kingdom (Q3, 1.13). Rank-size (in logarithms) plots in the upper panel correspond to sizes measured in demographic terms, while those in the lower panel reflect aggregate light emissions. The plots reveal a consistent concavity in the rank-size distributions, particularly for NTL-based measures, which becomes more pronounced as less urbanized settlements are included. This concavity underscores the concentration of population in the largest urban areas, with smaller settlements contributing to a flatter tail in the distribution. A similar but more pronounced pattern is observed for light emissions, where the upper ranks exhibit steeper slopes, highlighting the uneven distribution of economic activity. Notably, the degree of concavity and divergence between population- and light-based measures reflects the level of urbanization and economic development of each country. The United Kingdom exhibits the flattest

[8] Refer to Table A2 in the appendix for descriptive statistics.

[9] Country-specific estimation results are available from the authors upon request.

Table 2
Potential determinants of the city size distribution at the country level: Description of variables and data sources.

Variable	Description	Source
popul	Total population, persons	World Development Indicators
popgr	Annual population growth rate, percent	World Development Indicators
urban	Urban population, as percentage of total population	World Development Indicators
netmigr	Net migration, persons	World Development Indicators
ethnic	Ethnic fractionalization	Alesina et al. (2003)
rugged	Terrain ruggedness	Nunn and Puga (2012)
coastprox	Coastal proximity	Nunn and Puga (2012)
coastbord	Coastal border, kilometers	CIA World Factbook
area	Land area, square kilometers	World Development Indicators
latitude	Latitude	Nunn and Puga (2012)
malaria	Incidence of malaria, per thousand persons at risk	World Development Indicators
extreme	Droughts, floods, and extreme temperatures; as percentage of total population	World Development Indicators
colherit	Colonial heritage	CEPII GeoDist
govexp	General government final consumption expenditure, as percentage of GDP	World Development Indicators
democracy	Polity score	Center for Systemic Peace
intwar	Interstate war	Issue Correlates of War Project
indep	Time of independence	Issue Correlates of War Project
trade	Trade of goods and services, as percentage of GDP	World Development Indicators
resrents	Total natural resource rents, as percentage of GDP	World Development Indicators
gdp	Gross domestic product, in 2015 US Dollars	World Development Indicators
gdppc	Gross domestic product per capita, in 2015 US Dollars	World Development Indicators
manuf	Manufacturing, as percentage of GDP	World Development Indicators
services	Services, as percentage of GDP	World Development Indicators

slopes and less pronounced inequality, consistent with its advanced stage of development and mature urban system. In contrast, Malaysia and the Philippines display steeper slopes, indicative of more concentrated urban hierarchies typical of developing economies. Furthermore, the inclusion of smaller settlements results in a gradual flattening of the rank-size curves, which reduces the disparity between population- and light-emissions distributions, as shown in Fig. 3.

4. Potential determinants of national city size distributions

In line with the related literature, our main objective is to examine the variables that exhibit a robust relationship with the estimated slope parameters from national rank-size regressions. These coefficients are considered a measure of the unevenness in the size distribution of cities at the country level. We take the work of Düben and Krause (2021) as our starting point, as they compile an extensive list of potential determinants of national city size distributions. To avoid multicollinearity problems, we have excluded variables with correlation coefficients higher than 0.70. Following this approach, we have selected 26 covariates, that can be grouped into five categories: demography, geography, institutions, openness, and economy. Their descriptions and sources[10] are reported[11] in Table 2.

The majority of studies on the distribution of city sizes measure them in terms of population. This explains the consideration of demographic factors to account for the cross-country variation in estimated rank-size coefficients. For example, national and urban total populations, aside from being the most frequently employed control variables (Düben & Krause, 2021; Ioannides, Overman, Rossi-Hansberg, & Schmidheiny, 2008; Modica, 2017; Rosen & Resnick, 1980; Soo, 2005; Sun et al., 2021; Wang et al., 2021), have been found to be directly related to the equality displayed by national city size distributions. Conversely, the link with population growth and, especially, migration is ambiguous. This is because, on the one hand, Wang, Zhou, and Sun (2022) have shown that population growth does not have a statistically significant relationship with Pareto coefficients, and on the other hand, migrants usually head to densely populated cities. We have also considered the measure of ethnic fractionalization proposed by Alesina, Devleeschauwer, Easterly, Kurlat, and Wacziarg (2003) as it can influence agglomeration economies through economic growth, well-being, and institutional quality. Nonetheless, the magnitude of these effects is difficult to assess when these covariates are considered alongside demographic indicators.

Among the variables capturing geographical factors, land area has typically been highlighted since the pioneering work of Rosen and Resnick (1980). The underlying premise posits that nations with larger territories may have greater opportunities for spatial expansion, potentially resulting in a more concentrated distribution of cities; see Henderson and Wang (2007), Ioannides et al. (2008), and Modica (2017). Additionally, Wang et al. (2021) incorporate population density into their empirical analysis because this variable is likely directly related to the equality of the size distribution of cities at the country level. We have also considered other geographical factors as covariates, such as coastal proximity, borders, latitude, terrain ruggedness, incidence of malaria, and climate. Düben and Krause (2021) find an association between coastal proximity and a more unequal distribution of city sizes, and suggest that terrain ruggedness and latitude should be accounted for as factors influencing intra-country connectedness. They argue that these variables offer advantages over alternative indicators like roads or railways due to their inherent resistance to human modifications. Indeed, transport costs rise in complex terrains, contributing to a more uneven distribution of population (Soo, 2005). Moreover, Castells-Quintana, Krause, and McDermott (2021) suggest that climate change will likely impact the pace and nature of urbanization, potentially affecting spatial development patterns and overall welfare. Accordingly, we have considered a variable reflecting the incidence of droughts, floods, and extreme temperatures.

We have incorporated several institutional factors into our analysis, including colonial heritage, government consumption, the Polity score, indicators of intrastate war, and the date of independence. These variables serve as proxies for attributes of liberty and/or human rights, and we expect them to directly influence the estimated rank-size regression coefficients. This prediction is based on the premise that greater stability within a country enables the allocation of more funds and resources to smaller urban centers, thus enhancing their appeal and fostering a more balanced size distribution (Henderson & Wang, 2007; Sun et al., 2021; Wang et al., 2021). The size of government can be interpreted as a marker of both dictatorship and stability (Soo, 2005). In the context of dictatorship, a higher proportion of government consumption relative to the economy might diminish market forces, leading elites to concentrate

[10] Missing values in the original sources have been filled using alternative data sets, primarily the Economic Indicators provided by Moody's Analytics (https://www.economy.com/indicators).

[11] Table A3 in the appendix shows descriptive statistics for the potential determinants of national city size distributions considered in our analysis.

in national capitals or larger cities to gain better access to public services and resources (Ioannides et al., 2008). Conversely, increased government expenditure can mitigate regional inequalities through the redistribution of tax revenues (Modica, 2017). In terms of the urban impact of conflicts, Glaeser and Shapiro (2002) found that large cities are both highly prized targets and the most protected areas by national governments during wars. Depending on which effect is more dominant, this will result in either a higher or lower centralization of the urban structure, respectively.

Trade in goods and services as a percentage of GDP has been considered as a potential determinant of the equality in the distribution of city sizes, despite its ambiguous influence (Modica, 2017; Soo, 2005; Sun et al., 2021). According to Wang et al. (2021), this ambiguity can be attributed to a non-linear relationship between international trade and the spatial distribution of population and economic activity. Initially, globalization allows larger cities to gain more benefits from increased trade, leading to greater population polarization. However, over time, smaller cities may develop opportunities that contribute to a more equitable size distribution. Additionally, we have incorporated several variables reflecting national economic conditions and structure: GDP, GDP per capita, the shares of the manufacturing and services sectors in GDP, as well as total rents from natural resources. It is predicted that countries with higher levels of development will tend to exhibit more uniform city size distributions.

5. Bayesian model averaging

The broad array of potential factors influencing cross-country variation in city size distributions underscores the importance of explicitly accounting for model uncertainty in empirical analyses within this context. As noted by Fernández, Ley, and Steel (2001), failing to account for the uncertainty associated with the choice of covariates in linear regression models can lead to overestimating the precision of obtained results and, consequently, to an overly confident interpretation of the importance of specific predictors. Model averaging in a Bayesian framework is a practical and effective tool for addressing model uncertainty, particularly in empirical contexts characterized by a large number of potential models and relatively limited observations (Steel, 2016), as in our case. This approach facilitates the evaluation of multiple regressors by estimating all possible models and computing a weighted average of their results. In doing so, it accounts for the uncertainty inherent both within individual models and across different models. Fundamentally, BMA combines model selection, estimation, and inference into a cohesive framework, offering a robust and comprehensive empirical approach.

Assuming that the estimated coefficient from a rank-size regression linearly depends on a vector of covariates x, its conditional mean is given by:

$$E(\widehat{\beta}_c|x_c) = x_c'\theta, \quad c = 1,\ldots,C; \tag{2}$$

where C is the number of countries and θ is a set of parameters, estimated using maximum likelihood.

Model uncertainty is related to the choice of regressors to include in x. More specifically, given a total of q variables, there are 2^q models (sets of regressors) to be estimated, denoted as M_j for $j = 1,\ldots,2^q$. Each model M_j depends on a set of parameters θ^j with the conditional posterior probability:

$$g(\theta^j|\widehat{\beta}, M_j) = \frac{f(\widehat{\beta}|\theta^j, M_j)g(\theta^j|M_j)}{f(\widehat{\beta}|M_j)}; \tag{3}$$

where $f(\widehat{\beta}|\theta^j, M_j)$ and $g(\theta^j|M_j)$ denote, respectively, the likelihood function and the prior.

Given a prior model probability $P(M_j)$, the posterior probability can be calculated using Bayes' rule:

Table 3
Determinants of the size distribution of urban centers at the country level: Bayesian model averaging.

Variable	Population			Light emissions		
	PIP	Mean	SD	PIP	Mean	SD
popul	0.37	−3.72E−12	5.46E-11	0.28	1.05E-11	5.78E-11
popgr	0.37	-1.48E-05	0.01	0.37	−0.01	0.01
urban	0.65	−9.53E-04	1.04E-03	0.63	1.16E-03	1.22E-03
netmigr	0.39	3.64E-09	1.19E-08	0.36	7.70E-09	1.72E-08
ethnic	0.37	0.01	0.04	0.29	−0.01	0.04
rugged	0.36	2.29E-04	0.01	0.43	−0.01	0.01
coastprox	0.38	−0.01	0.03	0.25	1.70E-04	0.02
coastbord	0.43	−2.28E-07	5.07E-07	0.26	−5.81E-08	3.74E-07
area	0.37	1.51E-10	4.50E-09	0.27	−5.02E-10	4.15E-09
malaria	0.39	−2.62E-05	9.92E-05	0.29	−2.32E-05	9.85E-05
extreme	0.60	0.01	0.01	0.50	−0.01	0.01
resrents	0.42	−6.85E-04	1.79E-03	0.91	−0.01	3.22E-03
latitude	0.86	1.41E-03	8.86E-04	0.28	1.11E-04	4.65E-04
colherit	0.44	−0.02	0.04	0.38	0.02	0.04
govexp	0.37	2.57E-04	1.84E-03	0.53	2.68E-03	3.44E-03
democracy	0.39	−8.45E-04	2.70E-03	0.41	2.11E-03	3.80E-03
intwar	0.40	0.01	0.02	0.25	1.39E-03	0.02
indep	0.63	−6.34E-05	7.38E-05	0.73	−1.05E-04	8.73E-05
trade	0.47	−1.90E-04	3.64E-04	0.82	8.63E-04	5.83E-04
gdp	0.41	−1.99E-09	5.70E-09	0.29	−1.51E-09	5.87E-09
gdppc	0.38	1.44E-07	8.81E-07	0.34	4.26E-07	1.04E-06
manuf	0.41	5.94E-04	1.63E-03	0.35	8.55E-04	1.91E-03
services	0.88	−3.67E-03	2.20E-03	0.28	2.02E-04	1.15E-03
Models	1,817,347			1,444,514		
Size	10.72			9.50		
Correlation	0.85			0.98		
Shrinkage	0.68			0.89		

Note: The dependent variable is the estimated slope parameter from a rank-size OLS regression at the country level. The number of observations is 94. The birth-death MC3 sampler has been implemented with 500,000 burn-ins and two million iteration draws. The hyper-g and uniform priors have been established, respectively, for parameters and models. PIP denotes the posterior inclusion probability of each variable. Mean and SD are the posterior mean and standard deviation from model averaging. The lower panel reports the number of models visited, their average size, the correlation between iteration counts and analytical posterior model probabilities, and the mean of the shrinkage factor.

$$P(M_j|\widehat{\beta}) = \frac{f(\widehat{\beta}|M_j)P(M_j)}{f(\widehat{\beta})} \tag{4}$$

Expressions (3) and (4) indicate the necessity of specifying priors, which are updated according to the data, for both model parameters and probabilities. Leamer (1978) assumed that θ is a function of θ^j to derive the posterior density function of parameters across all candidate models using the law of total probability. Posterior inclusion probabilities (PIP) for the q regressors can be calculated by aggregating the posterior probabilities of models that include them. Notably, Steel (2020) highlights these PIPs and model probabilities as advantages of BMA methodology. To avoid estimating the entire set of 2^q models, the BMS R package developed by Zeugner and Feldkircher (2015) employs a Metropolis-coupled Markov-chain Monte Carlo (MC3) sampler. After

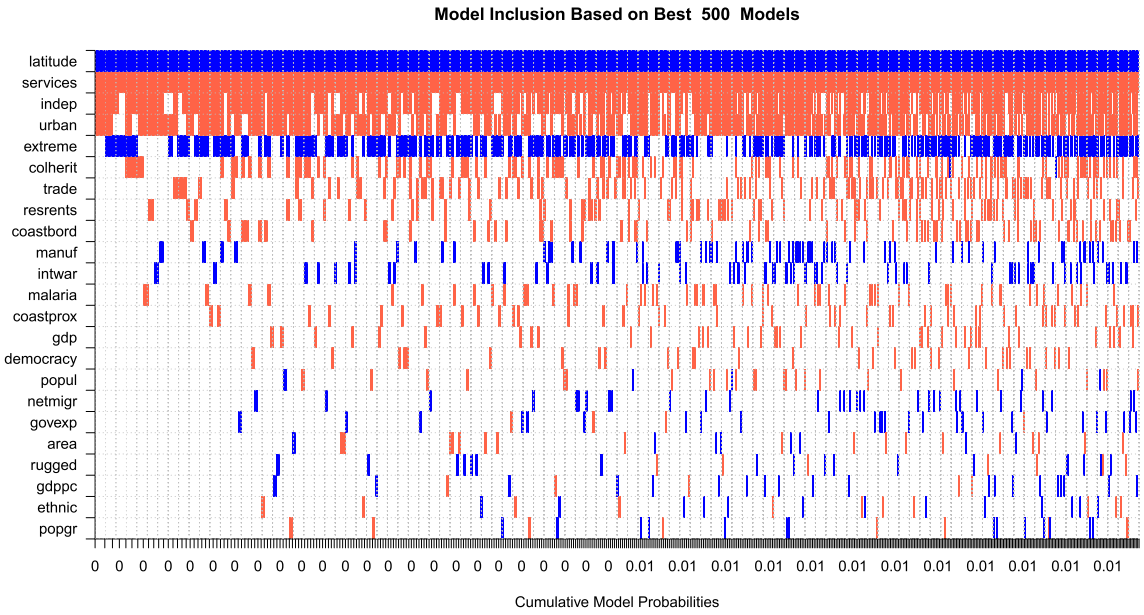


Fig. 5. Population in urban centers: MC3 sampler results of the 500 best models. Colored areas reflect the inclusion of variables in the model, and whether their estimated parameters are positive (blue) or negative (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

discarding the initial 500,000 draws ('burn-ins') to ensure convergence to an appropriate distribution, our empirical analysis considers two million subsequent iterations. We utilize a hyper-g prior for model-specific parameters and a uniform prior over the model space as baseline specifications.

6. Results

The first three columns of results in Table 3 present the PIP, mean, and standard deviation (SD) of the estimated parameters for each covariate when urban size is calculated using GHS-POP data. The inclusion probabilities indicate the importance of the variables in explaining the data, while the mean and standard deviation represent, respectively, a BMA point estimate and standard error. Consistent with the related literature, the results suggest that national city size distributions are influenced by economic, geographic, and institutional factors. Specifically, the share of services over GDP and latitude have inclusion probabilities exceeding 80 percent. The share of urban population, time of independence, and the relative occurrence of extreme climatic events also exhibit PIPs above 60 percent.

The figures reported in the lower panel of Table 3 indicate that the MC3 sampler has visited more than 1.8 million models, with an average size exceeding 10 covariates. The correlation coefficient between iteration counts and analytical posterior model probabilities for the 500 best models is 0.85, suggesting an adequate degree of convergence. Additionally, the average shrinkage factor across all models, which can

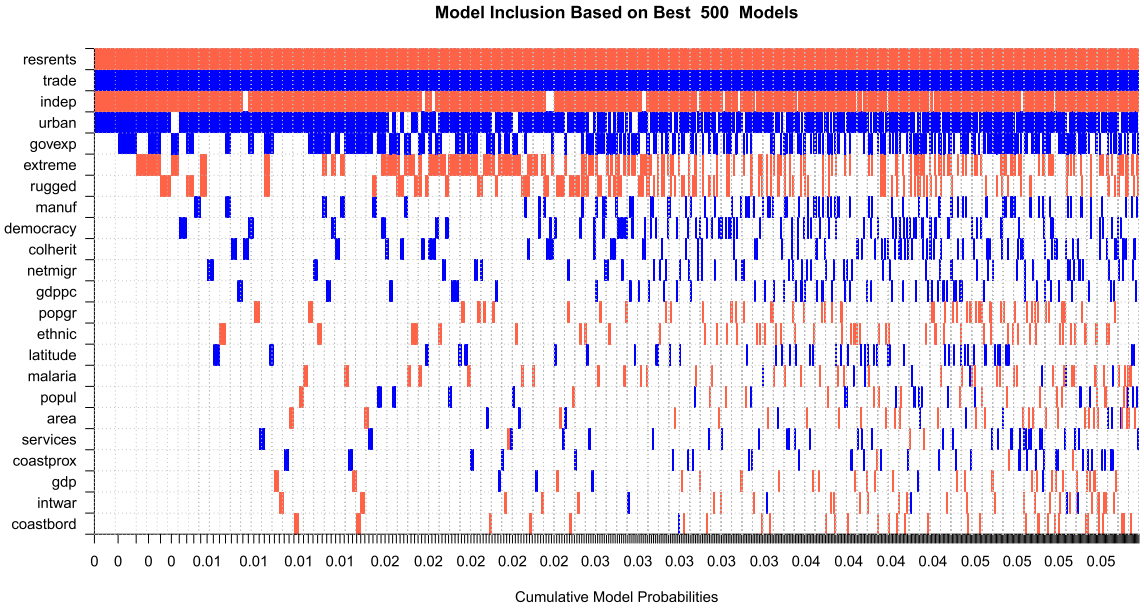


Fig. 6. Light emissions in urban centers: MC3 sampler results of the 500 best models. Colored areas reflect the inclusion of variables in the model, and whether their estimated parameters are positive (blue) or negative (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

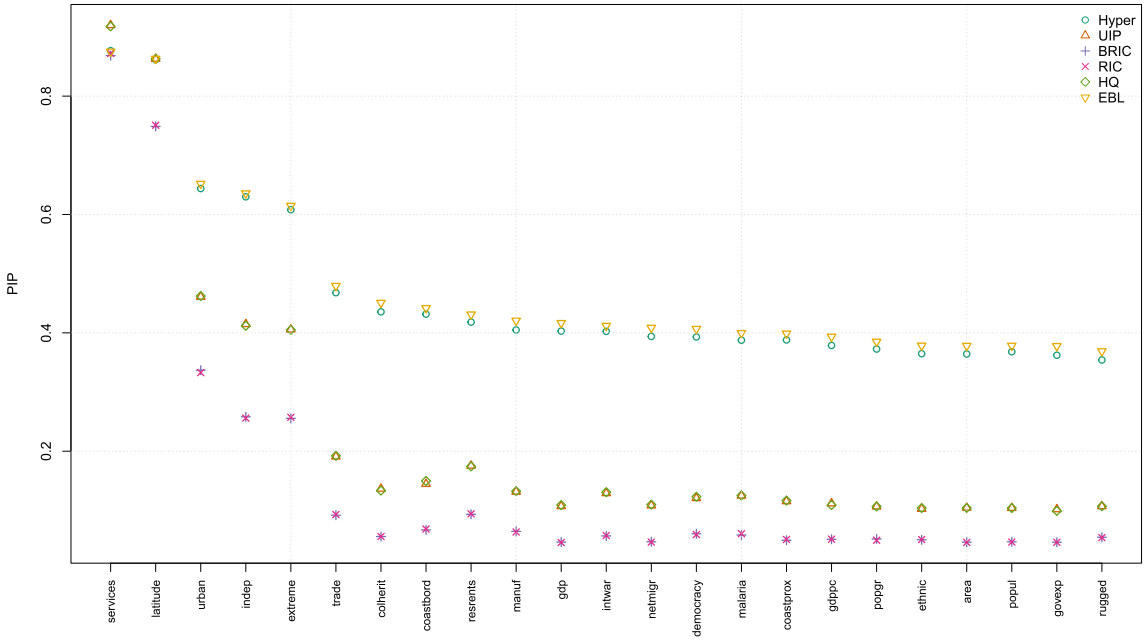


Fig. 7. Population in urban centers: Posterior inclusion probabilities. Sensitivity analysis to alternative specifications of the prior for model-specific parameters.

be interpreted as a Bayesian goodness-of-fit measure, is 0.68. The last three columns report the results for city sizes calculated using VIIRS data. The share of natural resource rents over GDP has the highest inclusion probability (91 percent), followed by the relative measure of trade volume of goods and services (82 percent). Similar to the results for city sizes in demographic terms, other variables with PIPs above 60 percent include the year of independence and the share of the population living in urban areas. The degree of convergence and the average shrinkage factor using NTL are even higher than those obtained when city size is measured using gridded population data.

A visual summary of the results described above is shown in Figs. 5 and 6 for urban population and light emissions, respectively. Each graph ranks the potential determinants of city size distribution vertically according to their PIPs. Horizontally, the best 500 models are ordered based on their posterior probabilities. A colored rectangle indicates the inclusion of a covariate in the model and shows the sign of its estimated influence (blue for positive, red for negative). Variables that tend to display high PIPs for both measures of city size are the time of independence, the share of urban population, and the incidence of extreme climatic events. Interestingly, while the time of independence shows an inverse relationship with the coefficient that proxies the equality of national city size distributions, the other two variables exhibit an opposite relationship with the distributions of urban population and light emissions. The urbanization rate is directly related to the equality of the national distribution of NTL but inversely related to that of the population. Conversely, the posterior mean parameters suggest an inverse relationship between the indicator of extreme climatic conditions and the equality of national city size distributions for NTL, while a direct

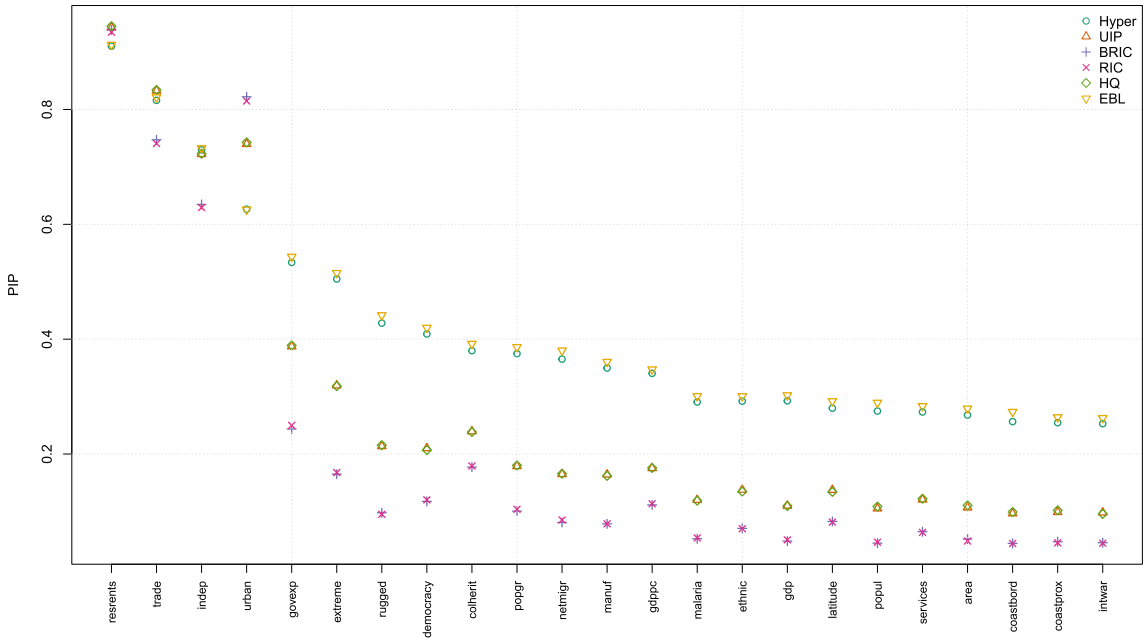


Fig. 8. Light emissions in urban centers: Posterior inclusion probabilities. Sensitivity analysis to alternative specifications of the prior for model-specific parameters.

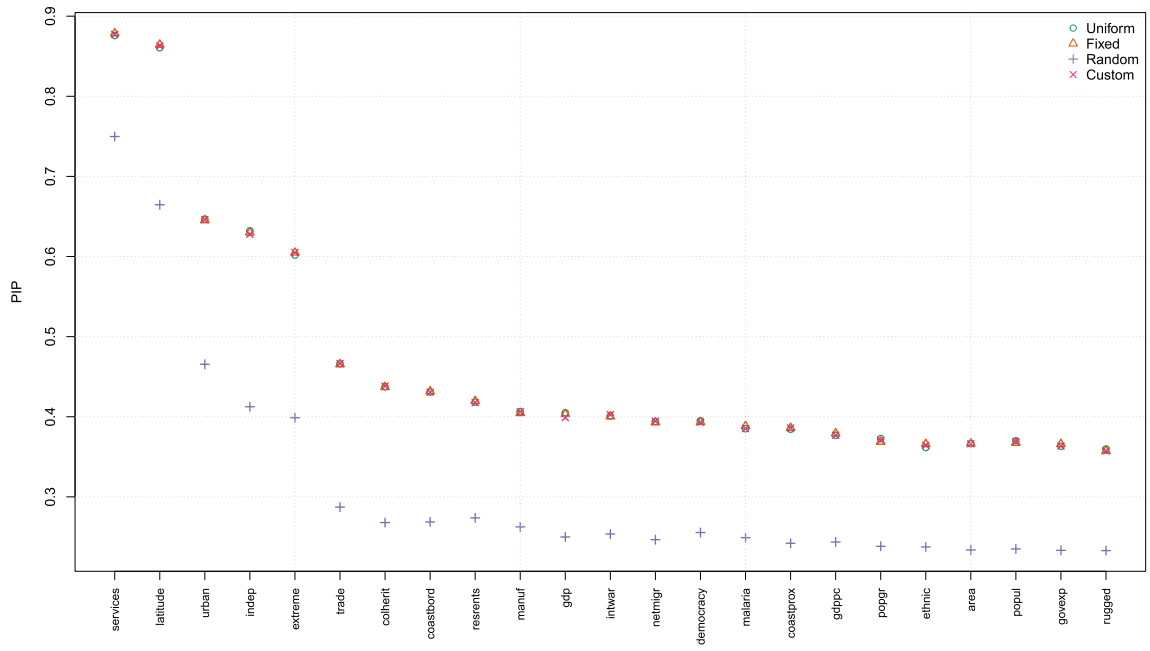


Fig. 9. Population in urban centers: Posterior inclusion probabilities. Sensitivity analysis to alternative specifications of model priors.

relationship is observed for the population. Therefore, it can be stated that the degree of urbanization and climate risk contribute to the observed differences between the distributions of population and light emissions in urban centers at the country level.

The choice of the prior for model-specific parameters may be influencing previous findings, as noted by Steel (2020). To assess their robustness, Figs. 7 and 8 present inclusion probabilities for the potential determinants of the parameters that characterize the distributions of urban population and light emissions, respectively, under different prior specifications; see Zeugner and Feldkircher (2015) and Forte, Garcia-Donato, and Steel (2018) for a description. It can be observed that the ranking of the variables according to their PIPs is minimally affected by the choice of the prior on model-specific parameters. With the exception of the local empirical Bayes prior (EBL), inclusion probabilities tend to be lower when constant g priors are used, particularly in the cases of the risk inflation criterion (RIC) and benchmark (BRIC) priors.

To evaluate the impact of the uniform model prior assumption, which assigns more probability mass to models of intermediate size, we considered three alternative specifications: (i) a fixed common prior inclusion probability for each regressor, such that the expected model size is $q/2$ (Fixed), (ii) a binomial-beta hyperprior on the prior inclusion probability (Random), and (iii) a custom inclusion probability of 0.5 (Custom). The results for each regressor under these model priors are presented in Figs. 9 and 10 for population and NTL, respectively. Notably, with the exception of the binomial-beta hyperprior, the inclusion probabilities under these alternative specifications are very

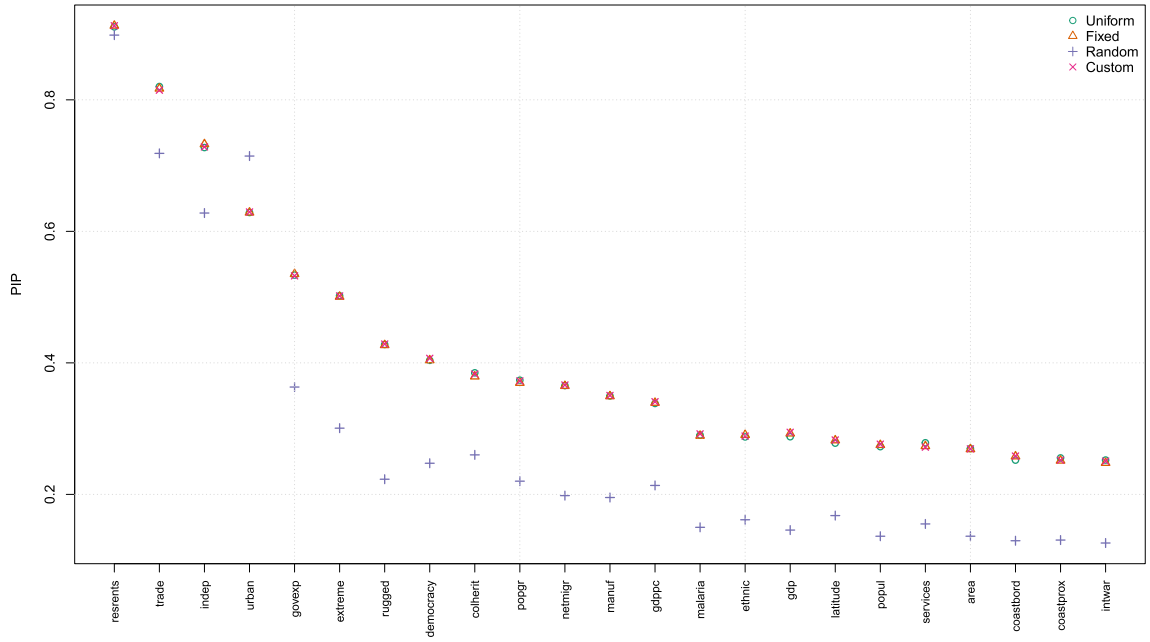


Fig. 10. Population in urban centers: Posterior inclusion probabilities. Sensitivity analysis to alternative specifications of model priors.

Table 4
Determinants of the city size distribution with high posterior inclusion probabilities: Alternative national sample compositions.

Population										
Urban centers			incl. Urban clusters			incl. Semi-dense towns			incl. Villages	
Variable	PIP	Sign	Variable	PIP	Sign	Variable	PIP	Sign	Variable	PIP	Sign
services	0.88	−	urban	0.71	−	urban	0.77	−	malaria	0.81	+
latitude	0.86	+	democracy	0.61	+	trade	0.63	+	democracy	0.69	+
urban	0.65	−	resrents	0.60	−	resrents	0.63	−	resrents	0.56	−
indep	0.63	−	malaria	0.58	+	democracy	0.57	+	trade	0.56	+
extreme	0.60	+	coastbord	0.56	−	malaria	0.53	+	coastprox	0.56	+
trade	0.47	−	latitude	0.55	+	coastprox	0.51	+	colherit	0.55	+
			govexp	0.54	+	latitude	0.50	+	popgr	0.55	+
			indep	0.52	−	govexp	0.50	+	rugged	0.51	−
			trade	0.49	+	colherit	0.49	+	latitude	0.48	+
Models	1,817,347		Models	2,108,870		Models	2,102,109		Models	2,046,407	
Size	10.72		Size	11.21		Size	11.30		Size	11.26	
Correlation	0.85		Correlation	0.35		Correlation	0.37		Correlation	0.42	
Shrinkage	0.68		Shrinkage	0.48		Shrinkage	0.50		Shrinkage	0.52	

Light emissions										
Urban centers			incl. Urban clusters			incl. Semi-dense towns			incl. Villages	
Variable	PIP	Sign	Variable	PIP	Sign	Variable	PIP	Sign	Variable	PIP	Sign
resrents	0.91	−	democracy	0.98	+	democracy	0.92	+	latitude	0.78	+
trade	0.82	+	popgr	0.67	−	popgr	0.79	−	democracy	0.62	+
indep	0.73	−	gdppc	0.59	+	gdppc	0.65	+	popgr	0.60	−
urban	0.63	+	resrents	0.53	−	trade	0.61	+	coastprox	0.49	−
govexp	0.53	+	trade	0.49	+	netmigr	0.59	+	extreme	0.48	−
extreme	0.50	−				resrents	0.53	−			
rugged	0.43	−				latitude	0.48	+			
Models	1,444,514		Models	1,564,321		Models	1,448,348		Models	1,934,326	
Size	9.50		Size	9.60		Size	9.61		Size	10.04	
Correlation	0.98		Correlation	0.91		Correlation	0.95		Correlation	0.82	
Shrinkage	0.89		Shrinkage	0.89		Shrinkage	0.91		Shrinkage	0.63	

Note: See Table 3.

similar to those obtained with the uniform model prior. In summary, the robustness checks indicate that the conclusions regarding the variables exhibiting a stronger relationship with the estimated rank-size coefficients at the country level remain largely unaffected by changes in parameter specifications and model priors.

6.1. Alternative national sample compositions

As shown in Section 3, following the seminal contribution by Eeckhout (2004), estimated coefficients from rank-size regressions – our dependent variable in the BMA framework – depend heavily on the number and type of settlements considered. For this reason, we repeated the procedure, sequentially including smaller units with a less urban nature in the sample. The covariates that receive high inclusion probabilities, as well as the sign of their posterior mean estimated coefficients, are reported in Table 4. The results displayed in the upper panel refer to the national distributions of sizes calculated using gridded population, while those in the lower panel are based on the distributions derived from VIIRS data.

Latitude is positively related to the equality of national population distributions, although its inclusion probability decreases as more settlements are included in the sample. By proceeding this way, geographical factors related to coastal borders and proximity, malaria incidence, and the availability of natural resources become relevant in explaining cross-country variation in the spatial distribution of population. Moreover, our results suggest that the evenness of population distribution is inversely related to the urbanization rate when villages are excluded from the rank-size regression estimations. The volume of trade as a percentage of GDP and, except in urban centers, the Polity score are directly related to the rank-size coefficients estimated from GHS-POP data at the country level.

The democratic nature of political regimes also displays high inclusion probabilities when size is measured using light emissions and as more units are included in the sample. On average, coefficients for the Polity score also present a positive sign, except when villages are considered. Rents from natural resources and international trade consistently show strong relationships with the equality in the distribution of economic activity across units. Additionally, while there is an inverse relationship between rank-size coefficients and revenues from natural resources, the opposite holds true for the volume of trade as a percentage of GDP. Developed countries, characterized by higher GDP per capita and lower population growth rates, are also expected to exhibit more uniform distributions of light emissions.

6.2. Comparison with standard linear regressions

Standard linear regression relies on a single model with a fixed set of variables, which may underestimate the uncertainty associated with model selection. Additionally, choosing a specific set of covariates can introduce bias if important variables are omitted or irrelevant ones included. In contrast, BMA accounts for a range of models, each weighted according to its associated probability. By considering all possible subsets of regressors, BMA reduces the risk of excluding relevant variables or including unnecessary ones. Furthermore, BMA incorporates an automatic model selection process by calculating posterior probabilities for each model, thus eliminating the need for subjective decisions about which variables to include. By explicitly accounting for model uncertainty, BMA is expected to provide more robust and realistic estimates. This is particularly crucial in our context, where the number of regressors is large relative to the cross-sectional dimension.

For comparison purposes, and to highlight the benefits of the empirical approach adopted in this study, Table 5 presents the OLS estimates using a demographic measure of city size. Although the adjusted

Table 5
Determinants of national city size distributions by population: OLS regressions.

Variable	Urban centers	+ Urban clusters	+ Semi-dense towns	+ Villages
popul	1.32E−11	−1.50E-03	-1.77E-10	-1.32E-10
	(1.41E−10)	(9.68E-03)	(1.18E-10)	(9.89E-11)
popgr	−0.01 (0.03)	0.01 (0.02)	0.01 (0.02)	−0.01 (0.02)
urban	−2.59E-03*	−1.70E-03*	−2.62E-03**	−5.60E-04
	(1.34E-03)	(9.21E-04)	(1.12E-03)	(9.41E-04)
netmigr	2.83E-08	−3.52E-08	−3.79E-08	−2.70E-08
	(3.13E-08)	(2.15E-08)	(2.61E-08)	(2.19E-08)
ethnic	0.10 (0.09)	0.02 (0.06)	0.04 (0.07)	0.02 (0.06)
rugged	−0.01 (0.02)	3.71E-03	−5.82E-04	−0.02 (0.01)
		(0.01)	(0.02)	
coastprox	−0.06 (0.06)	0.01 (0.04)	0.05 (0.05)	0.07 (0.04)
coastbord	−1.34E-06	−1.11E-06*	−9.29E-07	−2.98E-07
	(9.01E-07)	(6.18E-07)	(7.51E-07)	(6.31E-07)
area	9.25E-09	8.53E-09	6.61E-09	−2.20E-10
	(1.15E-08)	(7.91E-09)	(9.61E-09)	(8.08E-09)
malaria	−7.56E-05	1.32E-04	1.02E-04	2.95E-04*
	(2.14E-04)	(1.47E-04)	(1.78E-04)	(1.50E-04)
extreme	0.02 (0.01)	−3.64E-03	−0.01 (0.01)	−0.01 (0.01)
		(8.50E-03)		
resrents	−1.36E-03	−3.86E-03*	−5.06E-03*	−4.05E-03*
	(3.27E-03)	(2.25E-03)	(2.72E-03)	(2.29E-03)
latitude	2.96E-03***	9.55E-04	4.51E-04	2.22E-04
	(1.06E-03)	(7.24E-04)	(8.80E-04)	(7.40E-04)
colherit	−0.10 (0.06)	−0.01 (0.04)	0.03 (0.05)	0.07 (0.05)
govexp	1.81E-03	4.54E-03*	4.76E-03	1.24E-03
	(3.85E-03)	(2.64E-03)	(3.21E-03)	(2.70E-03)
democracy	−3.61E-03	5.88E-03*	6.33E-03	7.68E-03**
	(5.03E-03)	(3.45E-03)	(4.20E-03)	(3.53E-03)
intwar	0.03 (0.04)	2.51E-03	3.22E-03	0.04 (0.03)
		(0.03)	(0.03)	
indep	−1.58E-04*	−8.80E-05	−7.21E-05	−2.94E-05
	(9.25E-05)	(6.34E-05)	(7.71E-05)	(6.48E-05)
trade	−8.79E-04	4.64E-04	9.08E-04	5.67E-04
	(5.78E-04)	(3.97E-04)	(4.82E-04)	(4.06E-04)
gdp	−1.97E-08	6.81E-09	1.31E-08	8.19E-09
	(1.35E-08)	(9.25E-09)	(1.12E-08)	(9.45E-09)
gdppc	2.10E-06	3.23E-08	5.65E-07	−4.75E-07
	(1.80E-06)	(1.24E-06)	(1.50E-06)	(1.26E-06)
manuf	2.94E-03	8.96E-04	1.71E-03	4.07E-04
	(3.02E-03)	(2.07E-03)	(2.51E-03)	(2.11E-03)
services	−6.02E-03**	−9.18E-04	−8.31E-04	−1.06E-03
	(2.50E-03)	(1.71E-03)	(2.08E-03)	(1.75E-03)
intercept	1.73*** (0.25)	1.02***	1.05***	0.75***
		(0.17)	(0.21)	(0.17)
RSE	0.15	0.10	0.12	0.10
Adj. R^2	0.21	0.09	0.11	0.12

Note: Standard errors reported in parentheses. RSE denotes residual standard error.
*** $p < 0.01$.
** $p < 0.05$, and
* $p < 0.10$.

Table 6
Determinants of national city size distributions by light emissions: OLS regressions.

Variable	Urban centers	+ Urban clusters	+ Semi-dense towns	+ Villages
popul	1.42E-10	3.06E-11	6.45E-11	7.50E-11
	(1.35E-10)	(8.45E-11)	(9.54E-11)	(7.17E-11)
popgr	−0.03 (0.03)	−0.02 (0.01)	−0.04* (0.02)	−0.02 (0.01)
urban	5.77E-04	−9.93E-04	−6.86E-04	−4.86E-05
	(1.28E-03)	(8.04E-04)	(9.08E-04)	(6.83E-04)
netmigr	3.48E-08	1.94E-08	3.06E-08	1.04E-08
	(3.00E-08)	(1.87E-08)	(2.12E-08)	(1.59E-08)
ethnic	−4.95E-03	−0.03 (0.05)	−0.02 (0.06)	0.03 (0.04)
	(0.08)			
rugged	−0.02 (0.02)	−0.01 (0.01)	−0.01 (0.01)	−0.01 (0.01)
coastprox	0.03 (0.05)	0.04 (0.03)	0.05 (0.04)	0.02 (0.03)
coastbord	−3.90E-07	−5.19E-07	−4.51E-07	1.31E-08
	(8.62E-07)	(5.40E-07)	(6.09E-07)	(4.58E-07)
area	−5.55E-09	−5.66E-09	−7.15E-09	−2.98E-09
	(1.10E-08)	(6.90E-09)	(7.79E-09)	(5.86E-09)
malaria	−7.26E-05	−1.60E-04	−1.42E-04	6.34E-05
	(2.05E-04)	(1.28E-04)	(1.45E-04)	(1.09E-04)
extreme	−0.01 (0.01)	−0.01 (0.01)	−0.01 (0.01)	−4.76E-03
				(0.01)
resrents	−6.98E-03**	−2.47E-03	−2.93E-03	−6.25E-04
	(3.13E-03)	(1.96E-03)	(2.21E-03)	(1.66E-03)
latitude	−3.06E-04	−1.24E-04	1.12E-04	1.01E-03*
	(1.01E-03)	(6.32E-04)	(7.13E-04)	(5.36E-04)
colherit	0.04 (0.06)	0.01 (0.04)	0.01 (0.04)	0.02 (0.03)
govexp	3.91E-03	1.84E-03	2.33E-03	−1.60E-03
	(3.68E-03)	(2.30E-03)	(2.60E-03)	(1.96E-03)
democracy	4.62E-03	0.01***	0.01**	3.33E-03
	(4.82E-03)	(3.02E-03)	(3.40E-03)	(2.56E-03)
intwar	0.01 (0.04)	−1.78E-03	1.96E-03	3.60E-03
		(0.02)	(0.03)	(0.02)
indep	−1.19E-04	−3.20E-05	−1.69E-05	7.59E-07
	(8.85E-05)	(5.54E-05)	(6.25E-05)	(4.70E-05)
trade	8.64E-04	3.39E-04	5.35E-04	−1.70E-04
	(5.54E-04)	(3.46E-04)	(3.91E-04)	(2.94E-04)
gdp	−1.31E-08	−8.54E-09	−1.02E-08	−8.13E-09
	(1.21E-08)	(8.07E-09)	(9.12E-09)	(6.85E-09)
gdppc	1.25E-06	1.75E-06	2-05E-06	6.73E-07
	(1.72E-06)	(1.08E-06)	(1.22E-06)	(9.16E-07)
manuf	2.09E-03	1.47E-03	1.52E-03	1.02E-04
	(2.89E-03)	(1.81E-03)	(2.00E-03)	(1.53E-03)
services	−4.28E-04	4.30E-04	6.89E-04	−3.80E-04
	(2.39E-03)	(1.50E-03)	(2.04E-03)	(1.27E-03)
intercept	0.86*** (0.24)	0.68***	0.63***	0.62***
		(0.15)	(0.17)	(0.13)
RSE	0.14	0.09	0.10	0.08
Adjusted R^2	0.43	0.47	0.52	0.10

Note: Standard errors reported in parentheses. RSE denotes residual standard error.
*** $p < 0.01$.
** $p < 0.05$, and
* $p < 0.10$.

R-squared is low, regressors that are statistically significant tend to receive inclusion probabilities above 54 %. Nonetheless, the standard linear regression framework understates the relevance of geographical factors – captured by latitude – in explaining cross-country variation in urban concentration, particularly when smaller settlements are included in the sample. A similar pattern is observed for openness to trade: despite consistently receiving PIPs above 47 % in the BMA framework, it never appears as significant in the OLS regressions.

Table 6 presents the OLS estimation results when city size is proxied by aggregate light emissions. As in the previous case, significant covariates also receive high inclusion probabilities under the BMA framework, which proves to be far more informative than standard linear regression. While its explanatory power is generally higher for this proxy of economic activity (with the exception of the sample that includes rural clusters), the number of statistically significant determinants of national size distributions remains limited. OLS estimates would have understated the negative influence of population growth and natural resource rents on the equality of the distribution of economic activity, particularly when smaller settlements are included. Moreover, the BMA approach makes it possible to identify a positive relationship between the Polity score – and especially openness to trade – and the estimated rank-size coefficients at the country level.

Using a linear regression framework, Soo (2005) argues that the estimated coefficient for trade openness in explaining cross-country variation in the concentration of city sizes – measured in demographic terms – was unstable and often statistically insignificant. Similarly, Modica (2017) reports that international trade exerts a (small) negative impact on the rank-size coefficient. These findings suggest that the relationship between openness to trade and the distribution of city sizes may be either masked by other covariates included in the models or inherently non-linear. As noted by Wang, Wei, and Sun (2022), openness initially benefits larger cities by reinforcing agglomeration effects;

however, as globalization deepens and technological diffusion improves, smaller cities begin to gain more from trade, thereby reducing overall urban polarization. Standard OLS regression models are likely to overlook such non-linearities. In contrast, the BMA approach – by averaging across multiple model specifications – and the inclusion of smaller settlements allow for a more nuanced assessment of the potential for trade openness to foster more balanced national urban structures.

7. Concluding remarks

The extensive body of research linking demographic, geographic, institutional, and economic variables to the degree of urban concentration, combined with the absence of a comprehensive theoretical framework, highlights the need for empirical assessments of city size distribution determinants that explicitly account for model uncertainty during inference. To address this gap, this paper employs Bayesian model averaging techniques to evaluate the impact of numerous potential factors on cross-country variability in population and nighttime light distributions – an essential step for forecasting future trends and informing long-term planning and sustainability initiatives. In this analysis, we adopt globally consistent definitions of human settlements and leverage recent high-resolution satellite imagery as a proxy for economic activity. Our findings reveal that openness to trade, natural resource rents, and the Polity score are robustly associated with the estimated coefficients from rank-size regressions at the national level.

When focusing exclusively on urban centers and measuring city size in demographic terms, the volume of trade in goods and services as a percentage of GDP exhibits an inverse relationship with the estimated rank-size coefficient. This contrasts with the predictions of economic geography models but is consistent with previous evidence reported by Soo (2005) and Modica (2017). However, broadening the scope to include smaller urban units reveals stronger support for a positive effect of openness on the equality of both population and light emission distributions at the country level. These observations reinforce the argument in the literature that increased international trade can weaken agglomeration forces (Fujita, Krugman, & Venables, 1999), thereby reshaping the spatial distribution of economic activity, reducing spatial inequality, and encouraging urban-rural relocations (Catão & Obstfeld, 2019).

Expanding the sample to include settlements beyond major urban centers also reveals a positive association between the Polity score and the equality of city size distributions. As argued by Henderson and Wang (2007), this may reflect the role of democratization in enabling regional representatives to promote a more equitable allocation of public resources across smaller cities, thus contributing to greater spatial balance within national urban systems. Similarly, a higher share of income from natural resources is linked to greater inequality in the size distribution, suggesting that the spatial allocation of property rights influences patterns of urban concentration (Dentinho, 2017). Together, these insights highlight the importance of institutional quality in supporting national urban structures that are more evenly distributed.

CRediT authorship contribution statement

Miguel Puente-Ajovín: Data curation, Writing – review & editing, Visualization, Software, Methodology, Formal analysis, Conceptualization. **Marcos Sanso-Navarro:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **María Vera-Cabello:** Data curation, Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Formal analysis, Conceptualization.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work the authors used ChatGPT in order to improve the readability and language of the manuscript. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the content of the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cities.2025.106184.

Data availability

Data will be made available on request.

References

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., & Wacziarg, R. (2003). Fractionalization. *Journal of Economic Growth, 8*(2), 155–194. https://doi.org/10.2307/40215942

An, G., Choi, N., Lee, J. H., Kim, M., & Lee, J. (2024). *A story of urban development in Korea: From overconcentration toward balanced territorial development and urban regeneration*. Washington, DC: World Bank Group. https://documents.worldbank.org/en/publication/documents-reports/documentdetail/099061824045029920/p17697411e0a8f0b91a72e1773502f30b22.

Appiah-Kubi, M. A., & Gyambibi, F. A. (2025). City size distribution of African countries: A spatial analysis using geospatial data. *Cities, 160*, Article 105808. https://doi.org/10.1016/j.cities.2025.105808

Castells-Quintana, D., Krause, M., & McDermott, T. K. J. (2021). The urbanizing force of global warming: The role of climate change in the spatial distribution of population. *Journal of Economic Geography, 21*(4), 531–556. https://doi.org/10.1093/jeg/lbaa030

Catão, L., & Obstfeld, M. (2019). *Policies to make trade work for all*. Princeton, NJ: Princeton University Press. https://doi.org/10.1515/9780691198866. ISBN: 9780691198866.

Chen, X., & Nordhaus, W. D. (2011). Using luminosity data as a proxy for economic statistics. *Proceedings of the National Academy of Sciences, 108*(21), 8589–8594. https://doi.org/10.1073/pnas.1017031108

Christiaensen, L., & Todo, Y. (2014). Poverty reduction during the rural-urban transformation – The role of the missing middle. *World Development, 63*, 43–58. https://doi.org/10.1016/j.worlddev.2013.10.002

Dentinho, T. P. (2017). Urban concentration and spatial allocation of rents from natural resources. A Zipf`s curve approach. *REGION, 4*(3), 77–86. https://doi.org/10.18335/region.v4i3.169

Düben, C., & Krause, M. (2021). Population, light, and the size distribution of cities. *Journal of Regional Science, 61*(1), 189–211. https://doi.org/10.1111/jors.12507

Duranton, G. (2015). Growing through cities in developing countries. *World Bank Research Observer, 30*(1), 39–73. https://doi.org/10.1093/wbro/lku006

Duranton, G., & Puga, D. (2014). "The growth of cities." In *Handbook of Economic Growth*, edited by. In P. Aghion, N. Steven, & Durlauf (Eds.), *2. Handbook of Economic Growth. Amsterdam, The Netherlands: Elsevier* (pp. 781–853). https://doi.org/10.1016/B978-0-444-53540-5.00005-7. ISBN: 9780444535467.

Eeckhout, J. (2004). Gibrat's law for (all) cities. *American Economic Review, 94*(5), 1429–1451. https://doi.org/10.1257/0002828043052303

Ester, M., Kriegel, H.-P., Sander, J., & Xiaowei, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the second international conference on knowledge discovery and data mining* (pp. 226–231). Portland, OR: AAAI Press. https://doi.org/10.5555/3001460.3001507.

Fernández, C., Ley, E., & Steel, M. F. J. (2001). Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics, 16*(5), 563–576. https://doi.org/10.1002/jae.623

Florczyk, A., Corbane, C., Ehrlich, D., Carneriro, S. M., Freire, T. K., Maffenini, L., Melchiorri, M., et al. (2019). *GHSL Data Package, 2019*. EUR 29788 EN. Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/290498.

Florczyk, A., Melchiorri, M., Corbane, C., Schiavina, M., Maffenini, L., Pesaresi, M., Politis, P., et al. (2019). *Description of the GHS Urban Centre database 2015* (JRC115586). Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/037310

Forte, A., Garcia-Donato, G., & Steel, M. (2018). Methods and tools for Bayesian variable selection and model averaging in normal linear regression. *International Statistical Review, 86*(2), 237–258. https://doi.org/10.1111/insr.12249

Fujita, M., Krugman, P., & Venables, A. J. (1999). *The spatial economy: Cities, regions, and international trade*. Cambridge, MA: The MIT Press. https://doi.org/10.7551/mitpress/6389.001.0001. ISBN: 9780262273329.

Gabaix, X. (1999). Zipf`s law for cities: An explanation. *Quarterly Journal of Economics, 114*(3), 739–767. https://doi.org/10.2307/2586883

Gabaix, X., & Ibragimov, R. (2011). Rank - 1/2: A simple way to improve the OLS estimation of tail exponents. *Journal of Business & Economic Statistics, 29*(1), 24–39. https://doi.org/10.1198/jbes.2009.06157

Glaeser, E. L., & Shapiro, J. M. (2002). Cities and warfare: The impact of terrorism on urban form. *Journal of Urban Economics, 51*(2), 205–224. https://doi.org/10.1006/juec.2001.2262

Henderson, J. V. (2003). The urbanization process and economic growth: The so-what question. *Journal of Economic Growth, 8*(1), 47–71. https://doi.org/10.1023/A:1022860800744

Henderson, J. V., Storeygard, A., & Weil, D. N. (2012). Measuring economic growth from outer space. *American Economic Review, 102*(2), 994–1028. https://doi.org/10.1257/aer.102.2.994

Henderson, J. V., & Wang, H. G. (2007). Urbanization and city growth: The role of institutions. *Regional Science and Urban Economics, 37*(3), 283–313. https://doi.org/10.1016/j.regsciurbeco.2006.11.008

Ioannides, Y. M., Overman, H. G., Rossi-Hansberg, E., & Schmidheiny, K. (2008). The effect of information and communication technologies on urban structure. *Economic Policy, 23*(54), 202–242. https://doi.org/10.1111/j.1468–0327.2008.00200.x

Krugman, P. (1991). Increasing returns and economic geography. *Journal of Political Economy, 99*(3), 483–499. https://doi.org/10.1086/261763

Leamer, E. E. (1978). *Specification searches: Ad hoc inference with nonexperimental data*. New York: John Wiley & Sons, Ltd.. ISBN: 9780471015208.

Modica, M. (2017). The impact of the European Union integration on the city size distribution of the member states. *Habitat International, 70*, 103–113. https://doi.org/10.1016/j.habitatint.2017.10.011

Nunn, N., & Puga, D. (2012). Ruggedness: The blessing of bad geography in Africa. *Review of Economics and Statistics, 94*(1), 20–36. https://doi.org/10.1162/REST_a_00161

Puente-Ajovin, M., Ramos, A., & Sanz-Gracia, F. (2020). Is there a universal parametric city size distribution? Empirical evidence for 70 countries. *Annals of Regional Science, 65*(3), 727–741. https://doi.org/10.1007/s00168-020-01001-6

Puente-Ajovin, M., Sanso-Navarro, M., & Vera-Cabello, M. (2022). The distribution of urban population and economic activity in the European Union and the United States. *Letters in Spatial and Resource Sciences, 15*, 517–522. https://doi.org/10.1007/s12076-022-00309-5

Puente-Ajovin, M., Sanso-Navarro, M., & Vera-Cabello, M. (2024). Comparing city size distributions: Gridded population versus nighttime lights. *Journal of Regional Science, 64*(4), 1323–1358. https://doi.org/10.1111/jors.12703

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association, 92*(437), 179–191. https://doi.org/10.2307/2291462

Rosen, K. T., & Resnick, M. (1980). The size distribution of cities: An examination of the Pareto law and primacy. *Journal of Urban Economics, 8*(2), 165–186. https://doi.org/10.1016/0094-1190(80)90043-1

Soo, K. T. (2005). Zipf`s law for cities: A cross-country investigation. *Regional Science and Urban Economics, 35*(3), 239–263. https://doi.org/10.1016/j.regsciurbeco.2004.04.004

Steel, M. F. J. (2016). Bayesian model averaging. In *Wiley StatsRef: Statistics reference online* (pp. 1–7). Hoboken, NJ: JohnWiley & Sons, Ltd.. https://doi.org/10.1002/9781118445112.stat07874

Steel, M. F. J. (2020). Model averaging and its use in economics. *Journal of Economic Literature, 58*(3), 644–719. https://doi.org/10.1257/jel.20191385

Sun, B., Zhang, T., Wang, Y., Zhang, L., & Li, W. (2021). Are megacities wrecking urban hierarchies? A cross-national study on the evolution of city-size distribution. *Cities, 108*, Article 102999. https://doi.org/10.1016/j.cities.2020.102999

Wang, Y., & Sun, B. (2024). The types of city size distributions and their evolution. *Cities, 150*, Article 105045. https://doi.org/10.1016/j.cities.2024.105045

Wang, Y., Sun, B., Li, S. W., .W., & Zhang, T. (2021). Can the internet reshape the national city size distribution? Cross-country evidence. *Papers in Regional Science, 100*(5), 1254–1272. https://doi.org/10.1111/pirs.12619

Wang, Y., Wei, Y. D., & Sun, B. (2022). New economy and national city size distribution. *Habitat International, 127*, Article 102632. https://doi.org/10.1016/j.habitatint.2022.102632

Wang, Y., Zhou, Y., & Sun, B. (2022). Equalization or polarization? The effect of the internet on national urban hierarchies across the world, 2000–2018. *Cities, 131*, Article 103989. https://doi.org/10.1016/j.cities.2022.103989

Wimberley, R., Morris, L., & Fulkerson, G. (2007). Mayday 23: World population becomes more urban than rural. *Rural Sociologist, 27*(1), 42–43.

World Bank. (2008). *World development report 2009: Reshaping economic geography*. Washington, DC: The World Bank. https://doi.org/10.1596/978-0-8213-7607-2. ISBN: 9780821376089.

Zeugner, S., & Feldkircher, M. (2015). Bayesian model averaging employing fixed and flexible priors: The BMS package for R. *Journal of Statistical Software, 68*(4), 1–37. https://doi.org/10.18637/jss.v068.i04