

Large cities lose their growth advantage as urban systems mature

Andrea Musso, Diego Rybski, Dirk Helbing, Frank Neffke

Papers in Evolutionary Economic Geography

25.34



Utrecht University

Human Geography and Planning

Large cities lose their growth advantage as urban systems mature

Andrea Musso^{1,2}, Diego Rybski^{2,3}, Dirk Helbing^{1,2}, Frank Neffke²

¹Computational Social Science, ETH Zurich, Zurich, 8006, Switzerland.

²Complexity Science Hub Vienna, Vienna, 1030, Austria.

³Leibniz Institute of Ecological Urban and Regional Development, Dresden, 01217, Germany.

Contributing authors: andrea.musso@gess.ethz.ch; ca-dr@rybski.de;
dirk.helbing@gess.ethz.ch; neffke@csh.ac.at;

Abstract

Abstract: The share of the world population living in cities with more than one million people rose from 11% in 1975 to 24% in 2025 (our estimates). Will this trend towards greater concentration in large cities continue or level off? We introduce two new city population datasets that use consistent city definitions across countries and over time. The first covers the world between 1975 and 2025, using satellite imagery. The second covers the U.S. between 1850 and 2020, using census microdata. We find that urban growth follows a characteristic life cycle. Early in urbanization, large cities grow faster than smaller ones. As urban systems mature, growth rates equalize across sizes. We use this life cycle to project future population concentration in large cities. Our projections suggest that 38% of the world population will be living in cities with more than one million people by 2100. This estimate is higher than the 33% implied by the well-known theory of proportional growth but lower than the 42% obtained by extrapolating current trends.

Keywords: Urbanization, Cities, Urban growth, Urban scaling

1 Introduction

Over the past 50 years, the world population has become increasingly concentrated in large cities. Whether this concentration will continue depends on the relationship between a city’s size and its growth rate, a topic on which the literature is divided.

Increasing returns theories, such as the New Economic Geography [1, 2] and the evolutionary approach of Pumain *et al.* [3–5], argue that large cities tend to grow faster than small ones. Large cities are typically more innovative and productive [6, 7], have more educated workforces [8, 9], sit more centrally in trade and knowledge networks [5], and host more advanced economic activities [10, 11]. These factors can plausibly confer them a growth advantage. Based on our novel projection model, even a mild growth advantage—where a 10-fold increase in size corresponds to a 0.7% increase in yearly growth—would result in 47% of the world population living in 1M+ cities by the end of the century.

Proportional growth theories argue instead that cities of all sizes tend to grow at similar rates (Gibrat’s Law) [12–16]. After all, scale also brings costs such as increased congestion [17, 18], crime [19, 20], disease risk [21, 22], and housing expenses [23, 24]; and these costs may offset advantages. If city size exhibits no growth advantage, our projections suggest that 33% of the world population will live in 1M+ cities in 2100.

This 14 percentage point gap between the two predictions represents a difference of 1.4 billion people and thus has significant implications for future planning and development. In this paper, we analyze which of these two scenarios is more likely using two newly built city population datasets. The first dataset, built from satellite-derived grids [25], spans 1975–2025 and covers 99 countries, together representing 94% of the world population. The second, built from census microdata [26–28], spans 1850–2020 and covers the United States (USA) urban system over nearly its entire history.

These datasets substantially extend the spatial and temporal coverage of previous efforts to build such harmonized data [29]. Before the advent of satellite imagery and geospatial processing tools, assembling harmonized datasets of city sizes across countries and years was extremely difficult, primarily because national data collection efforts are not designed to produce outputs that are comparable across countries or consistent over time. As a result, most empirical studies of urban growth focused on just a handful of countries over limited time periods. The conclusions of these studies are varied, some supporting proportional growth [13–16] and others not [30–35]. Integrating these results has proved difficult due to substantial differences in methodology [36], leaving us with an incomplete understanding of how urban growth evolves over time [37].

Using these new datasets, we sharpen our understanding of temporal trends in urban growth. First, we show that proportional growth and increasing returns are better understood as two phases of the same underlying process, not competing realities. In the early phases of urbanization, large cities enjoy a strong growth advantage, consistent with increasing returns. As an urban system matures, this advantage weakens, and the growth rates of large and small cities converge, consistent with proportional growth. Second, we show that this size-growth relationship translates into systematic changes in the shape of a country’s city-size distribution [38–40]. When large cities

Dataset	City-year obs.	Countries	Frequency (years)	Time Period
Global cities	1,604,593	99	5	1975 – 2025
USA cities	26,902	1	10	1850 – 2020

Table 1 Description of the two large-scale city population datasets created for the analysis. The Global cities dataset covers 99 countries around the world (amounting to 94% of the world population in 2025) between 1975 and 2025. Its primary source is the GHSL 2023 data package [25]. The USA cities dataset covers the continental USA between 1850 and 2020. Its primary sources are IPUMS USA [26], IPUMS NHGIS [27], and the Census Place Project [28].

have a growth advantage, the distribution stretches and its rank-size slope (a spline-based analogue of the Zipf exponent [41]) increases. Third, combining this mechanism and a novel projection method, we project that by 2100, 38% of the world’s population will live in 1M+ cities. This projection lies between the proportional-growth (33%) and increasing-returns (47%) benchmarks, and below an extrapolation of current trends (42%).

2 Results

Our city population datasets (Table 1) define “cities” geographically, as clusters of contiguous built-up areas or regions of high population density [33, 42, 43] (Methods 4.2). This definition has two main advantages: (i) it dynamically adjusts as urban areas expand; and (ii) it is consistent across space and time, facilitating robust comparative analyses.

Our analysis of these datasets reveals substantial global variation in urban growth patterns. Between 1975 and 2025, an average 1M+ city in Asia/Africa outgrew the national average by $\sim 7\%$ (Fig. 1A). In contrast, Europe’s large cities grew modestly faster than the rest (Fig. 1A-B), and in the Americas, city growth displayed an inverted-U-shaped trend, with 1M+ cities growing 1.6% slower than their national average (Fig. 1A-B).

These patterns are also visible at the country level. Figure 1C maps national size-growth slopes, β (definition in the caption of Figure 1C). A positive β indicates that growth increases with size, or, in other words, that large cities have a growth advantage. This growth advantage varies significantly across countries, with β ranging from -0.02 to 0.1 (median ≈ 0.012 ; Fig. 1C inset). Notably, β s are generally higher in Asia and Africa and lower in Europe and the Americas.

This regional divide can be understood as part of a universal dynamic in which the growth advantage of large cities weakens as a country’s urban system matures. This inverse relationship is evident in both cross-sectional and longitudinal data (Fig. 2).

Across countries, a lower national share of urban population correlates with higher β (Fig. 2A inset and Table 2). Put differently, growth rates increase rapidly with city size in less urbanized countries, whereas the size-growth relationship is nearly flat in more urbanized ones (Fig. 2A). The numbers speak clearly: between 1975 and 2025, 1M+ cities in more urbanized countries grew at the national average rate, while 1M+ cities in less urbanized ones grew 7.3% faster.

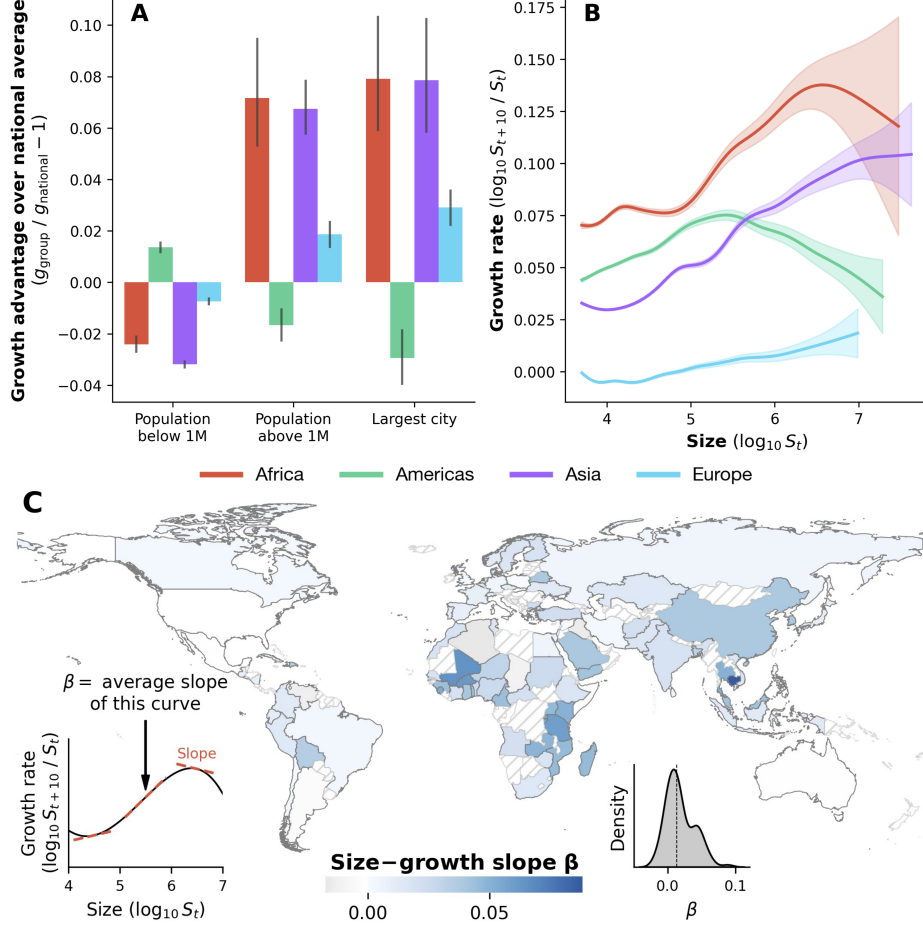


Fig. 1 The growth advantage of large cities varies systematically across regions: it is strong in Asia and Africa and weak in Europe and the Americas. **(A)** We compare the average growth rate of a country's cities, denoted by g_{national} , with the average growth rate of specific sub-groups of its cities (e.g., the largest city, 1M+ cities, etc.), denoted by g_{group} . Each bar in the chart shows the ratio $g_{\text{group}} / g_{\text{national}} - 1$ for a given region. **(B)** *Size-growth curve* by region. These curves are obtained by fitting a penalized cubic B-spline ($\lambda = 100$) to the relationship between city log-size in year t and city log-growth between year t and $t + 10$ (Methods 4.3.2). **(C)** *Size-growth slopes* β by country. β is obtained by first estimating the national size-growth curve (as in panel (B)), and then averaging the local slope of this curve across the size spectrum (see left inset and Methods 4.3.2). The map shows the mean value of β between 1975 and 2025; hatched indicates no data.

Within countries, β declines as urbanization rises. A regression with country fixed effects shows that a 20% increase in a country's urban population share is associated with a ~ 0.01 reduction in β (Table 2). This pattern is clearly observed in South Korea and the USA, two countries for which our data cover a large window of the urbanization process. Both countries saw large cities grow substantially faster than

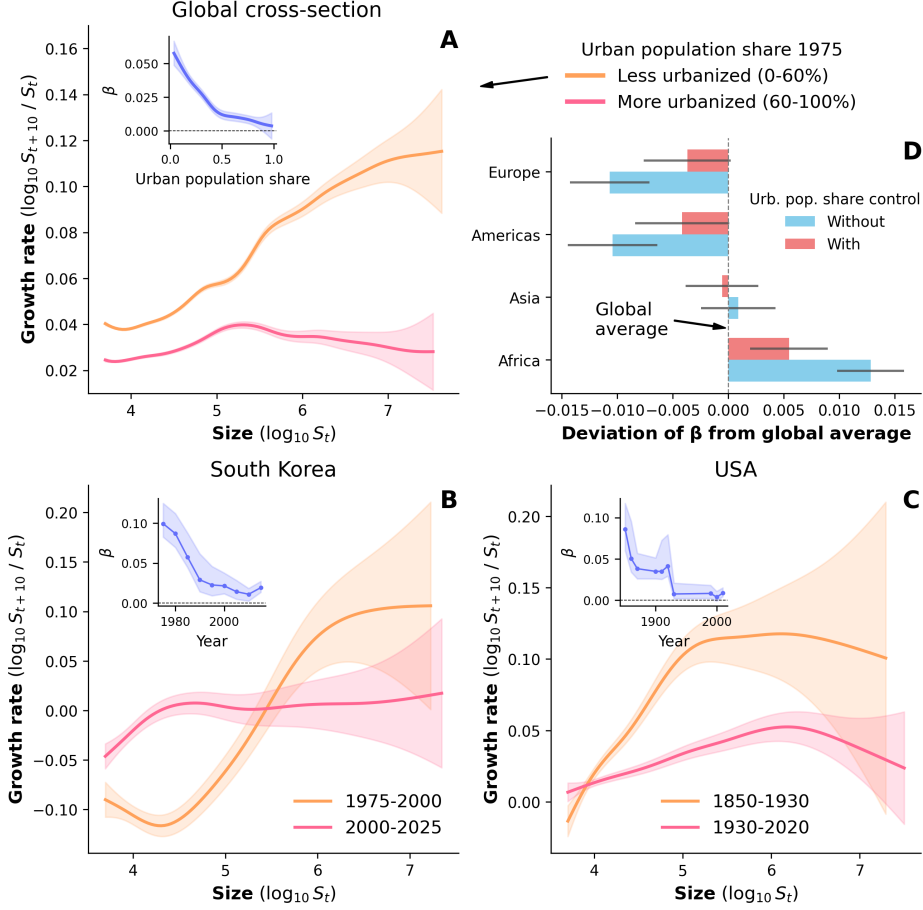


Fig. 2 The growth advantage of large cities weakens as countries urbanize. **(A)** Size-growth curve by urbanization group (Methods 4.3.2) (Inset) β vs. urban population share (fit with a penalized cubic B-spline with $\lambda = 100$). **(B-C)** Size-growth curves in South Korea and the USA across different time periods. (Insets) β vs. time (bootstrapped confidence intervals $n = 1000$). **(D)** Country-level regression with regional dummies ($\beta_{t,\text{country}} \sim \delta_{\text{region}} + \text{urban population share}_{t,\text{country}}$). Regional differences in β become significantly smaller once we control for the level of urbanization.

smaller ones in the early phases of urbanization, but by their mature phases, this advantage nearly vanished (Fig. 2B-C).

These results help explain the regional differences in β observed in Figure 1. A simple regression shows that when controlling for urbanization level, regional dummies converge toward the global average, with differences becoming either statistically insignificant or much smaller (Fig. 2D). This indicates that regional differences in β largely reflect each country's stage in the urbanization process.

The size-growth patterns observed in Figure 2 change national city size distributions. We can quantify this change by looking at the distribution's rank-size curve (Fig. 3A), or more precisely at the absolute value of its slope, α (exact definition of α

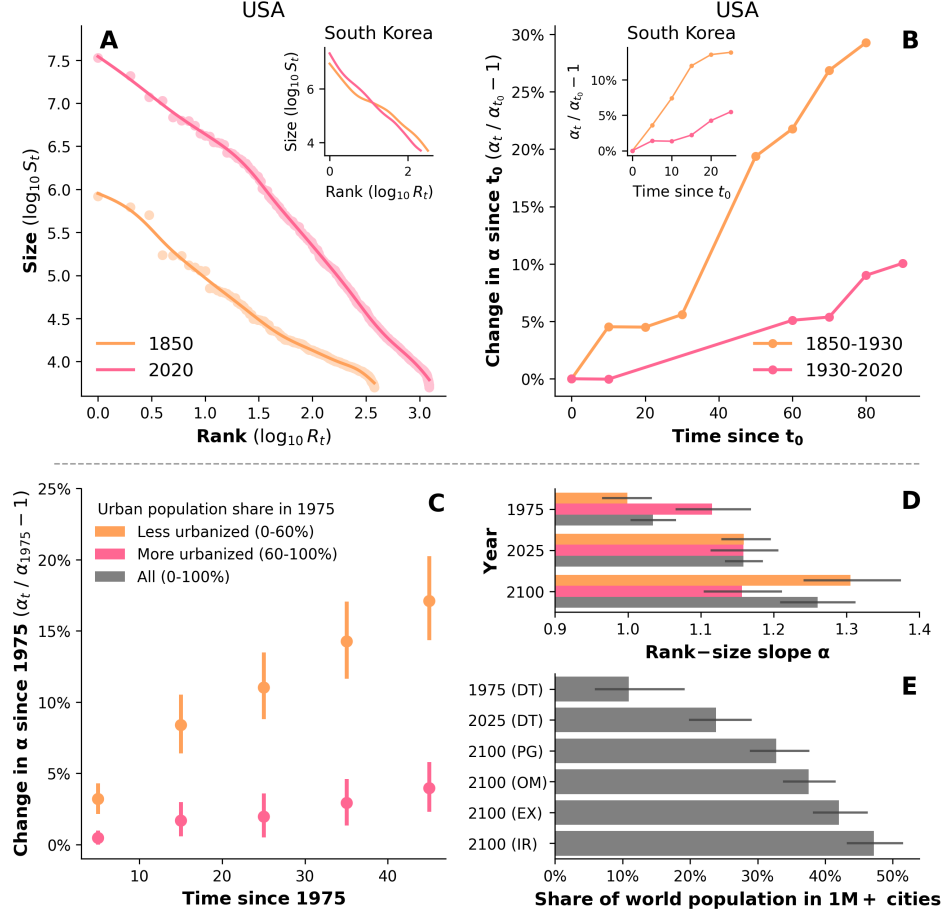


Fig. 3 Historical trends and projections for national city size distributions. **(A)** Rank-size curves for the USA and South Korea. These curves are obtained by fitting penalized cubic B-splines ($\lambda = 1$) to the relationship between city log-rank and city log-size in a fixed year. The USA plot displays also a scatter of the underlying data points. **(B-C)** The rank-size slope α is estimated by averaging the local slope of the rank-size curve across ranks and then taking the absolute value (Methods 4.3.2). Note that α is a spline-based analogue of the Zipf exponent. **(B)** Change in α for South Korea and the USA from a base year t_0 (South Korea $t_0 = 1975, 2000$; USA $t_0 = 1850, 1930$). **(C)** Change in average α across more/less urbanized countries since 1975 (binned scatter plot by year with bootstrapped confidence intervals; $n = 1000$). **(D)** Average α across groups of countries in different years. 2100 values are projected using the technique in Methods 4.4. **(E)** Share of the world population in 1M+ cities in 1975, 2025, and 2100. 2100 values are projected under four urban growth models, which differ in their estimates for the future trajectories of β : proportional growth (PG; $\beta = 0$), increasing returns (IR; $\beta = 0.03$), current trend extrapolation (EX; β equal to the country's average between 1975-2025), and our model (OM; time-varying β according to the country's urbanization). See Methods 4.4.

in the caption of Figure 3B-C). If β is positive, large cities have a growth advantage, the left tail of the rank-size curve rises faster than the right one, the curve steepens, and α increases. Thus, if β declines with urbanization (as in Fig. 2), α should increase at a decelerating rate. Our data confirm this: α increased in both the USA and South

Korea (Fig. 3B), and it did so faster early on. Further, since 1975, α grew 18% on average across less urbanized countries, against 4% across more urbanized ones (Fig. 3C).

With the mechanism in hand— β shapes α , and β weakens as countries urbanize—we move from description to prediction by projecting trajectories of national α s to 2100 (Methods 4.4). Our projections suggest that growth in α will slow over the coming decades. Between 1975 and 2025, α grew by 2.3% per decade, on average across the countries in our sample. Between 2025 and 2100 that same average is projected to grow at 0.9% per decade. Further, α will continue to grow faster in countries that were less urbanized in 1975. By 2100, their average α is projected to be 16% larger than that of the more urbanized group, while it was 12% smaller in 1975 (Fig. 3D).

Translating projected α s into population shares, we estimate that 38% of the world population will live in 1M+ cities by 2100 (Fig. 3E; Methods 4.4). This projection sits below an extrapolation of current trends (42%; Fig. 3E) and above the proportional-growth benchmark (33%; Fig. 3E), consistent with a weakening growth advantage of large cities as urban systems mature.

3 Discussion

Using a robust geographic definition of city and a comprehensive database spanning several countries and historical periods, we show that urban growth follows a common life cycle. Early in urbanization, large cities experience a strong growth advantage, stretching the city size distribution. As urban systems mature, this growth advantage weakens and the distribution stabilizes. We use this model to forecast urban concentration at the end of the century. Relative to an extrapolation of 1975-2025 trends, our model projects 450 million fewer residents in 1M+ cities by 2100 (−4.4%); relative to proportional growth, 490 million more (+4.8%).

This study has several limitations. First, despite the best harmonization efforts, the source data on which our analysis is based have shortcomings: GHSL grids undercount the population of rural areas [44], and 19th-century USA census records contain inconsistencies. Second, while we document trends in the relationship between city size and growth, we do not identify the mechanisms driving them. For example, we show that large cities grow faster in the early phases of urbanization, but we do not (i) model the imbalances in migration flows responsible for this pattern [37, 45]; or (ii) test the relative importance of different economic explanations, such as innovation diffusion [46] versus firm-level returns to scale [1]. Third, we focus on common size-growth trends, leaving cross-country variation around this trend underexplored. However, this variation is substantial, and what factors explain it — from national urban policies, industrial and governance structures to geography — is an important question.

Despite these limitations, our results have clear implications when viewed through the lens of urban scaling theory. Because many urban outcomes scale nonlinearly with population, reallocating people across cities of different sizes changes aggregate outcomes (see SI 2.3 and [47, 48]). For example, since productivity scales super-linearly, moving one person from a small city (10K) to a large one (1M) is associated with

productivity gains between 30-100%, depending on the estimate [6, 9]. Therefore, when large cities grow faster, reallocation acts as an engine of aggregate economic growth. As more urban systems mature and the growth advantage of large cities weakens, this engine loses power. As a result, reallocation-driven growth will slow, placing more of the burden on within-city productivity gains.

4 Methods

4.1 Data

This paper builds on several datasets, which are all publicly available. Below we provide a high-level overview of the main datasets. In Methods 4.5, we document how to retrieve them.

Population and Urbanization data: The population projections come from the United Nations World Population Prospect 2024 [49], the History database of the Global Environment 2023 [50], and Gapminder [51, 52], with processing from Our World in Data [53]. We use projections under the medium fertility scenario. The urbanization data come from Chen *et al.* [54]. We use their World Bank-based annual projections under SSP2 (“middle of the road”), which extend the World Bank series to 2100.¹

Country borders: Boundaries as of 2019, from the CShapes database [55].

Global grids: Our global cities dataset is based on the population (pop) and degree-of-urbanization (smod) grids from the Global Human Settlement Layer (GHSL) 2023 data package [25]. These grids estimate population counts and urbanization levels for 1km-by-1km cells covering the whole planet by combining satellite imagery with census data.

USA grids: Our USA cities dataset is based on custom grids derived from census place population estimates from various data sources. For the 1990-2020 period, we use estimates from the National Historical Geographic Information System (IPUMS NHGIS) [27]. For the 1850-1940 period, we reconstruct census place population estimates using the IPUMS USA full count census data [26] with geocoding from the Census Place Project [28]. To do so, we match over 500 million individual census records to approximately 40,000 census places. We then aggregate these records to estimate the population of each census place. Because this historical data comes from century-old handwritten documents, it contains numerous inconsistencies, such as census places disappearing and reappearing over time. We correct for these issues using several preprocessing steps that leverage individual migration data, as detailed in SI 1.1.

¹In principle, the urban population share could be computed directly from GHSL grids [25]. However, these grids are known to underestimate the rural populations [44], making such estimates unreliable. Therefore, we use GHSL grids only to compare urban areas, not to estimate the overall urban population share.

We use these clean census place population estimates to build population grids suitable for the City Clustering Algorithm. For each year t , we create a 1km-by-1km grid covering the continental US. Each grid cell c_j receives an initial population $\text{pop}_{\text{init}}(c_j)$, given by the sum of the population of all census places whose geographic center falls within that cell. We then smooth this initial grid using a spatial convolution kernel: the final population of cell c_i becomes a distance-weighted average of the initial population of neighboring cells c_j , with weights decreasing exponentially with distance $d_{c_i c_j}$:

$$\text{pop}(c_i) = \frac{1}{\sum_j e^{-\eta \cdot d_{c_i c_j}}} \sum_j e^{-\eta \cdot d_{c_i c_j}} \cdot \text{pop}_{\text{init}}(c_j) . \quad (1)$$

The decay parameter η , which controls how quickly population decays around a census place, is set to $\eta = 0.2$, similar to estimates in prior work [56].

4.2 Constructing cities: the City Clustering Algorithm

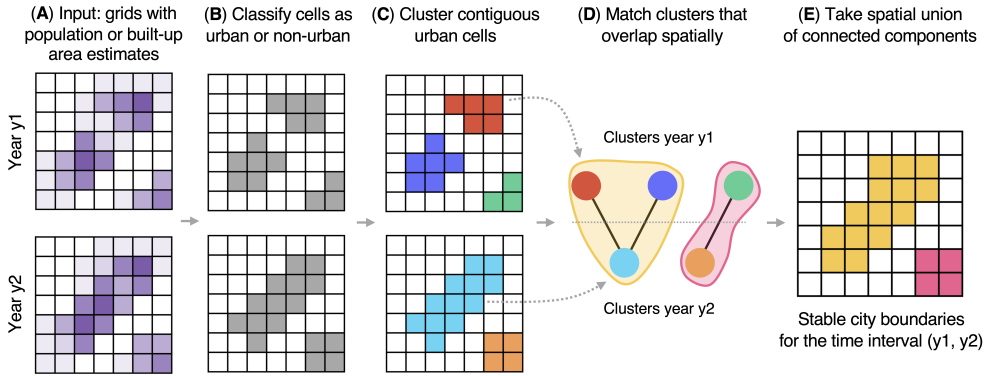


Fig. 4 The City Clustering Algorithm [33] computes stable city boundaries for a time interval (y_1, y_2) . This algorithm proceeds in five steps: **(A)** It takes two grids as input, one for each year, containing estimates of population or built-up area. **(B)** It classifies grid cells as either urban or non-urban using a threshold on these population/built-up area estimates. **(C)** It groups contiguous urban grid cells to form clusters. **(D)** It matches clusters across years when they overlap spatially, forming a bipartite graph. **(E)** It defines a city’s boundary as the spatial union of all clusters within a single connected component of the graph.

To construct cities from the USA and global grids, we use the City Clustering Algorithm [33]. This algorithm identifies stable city boundaries between two years t_1 and t_2 based on the city’s geographic extent, allowing for robust measurement of population growth and comparable city definitions across countries and times. The algorithm, illustrated in Figure 4, involves five steps:

1. **Classify urban cells:** For each year independently, we classify all grid cells as either “urban” or “non-urban” using population density or degree-of-urbanization thresholds.

2. **Form initial clusters:** For each year independently, we group contiguous *urban* cells into initial clusters using a flood fill algorithm [57].
3. **Match clusters over time:** For a pair of years $t_1 \leq t_2$, we construct a bipartite graph $G = (A, B)$, where nodes in part A are clusters in year t_1 , and nodes in part B are clusters in year t_2 . We connect clusters with an edge if and only if they overlap spatially.
4. **Define stable city boundaries:** We extract the connected components of the bipartite graph G that have at least one cluster in year t_1 ². Each component represents a single, evolving city that existed throughout the (t_1, t_2) period. The city’s boundary for the period is the spatial union of all clusters within the component (from both t_1 and t_2).
5. **Calculate population growth:** For each city, we calculate its population in years t_1 and t_2 by summing the population of all grid cells that fall within its boundary. The ratio of these two population values gives us the city’s growth rate over the period.

While the above algorithm largely follows the same approach as the original paper [33], we introduce some minor improvements and clarifications to its implementation.

First, using a bipartite graph to match clusters (Step 3) provides a more systematic and scalable approach than the manual specifications suggested in the original paper. It improves conceptual clarity, simplifies implementation, and naturally handles events like the merger or split of more than two clusters.

Second, our implementation is less ambiguous with respect to the “cell reclassification problem”. This problem occurs when cells flip their classification, moving from “non-urban” in year t_1 to “urban” in year t_2 (or vice versa). These flips do not require any dramatic transformation. As we use thresholds to determine “urban” and “non-urban” status, even small changes in population or built-up area can cause a flip. But if not treated carefully, flips may lead to biased growth estimates. For example, imagine a city where hundreds of grid cells flip from “non-urban” to “urban”. If we simply count the population in these newly “urban” cells as if they were entirely new additions to the city, we grossly overestimate its growth. Our solution (Steps 4 and 5) is to define a stable city boundary for the entire period and measure population changes only within that boundary. This isolates true population growth from the artifacts of reclassification. While this solution was adopted in the original CCA paper, the implementation is framed ambiguously and the problem is not mentioned.

Third, our implementation addresses a form of selection bias that we call “new cluster bias”. This bias arises if one includes urban clusters that appear in year t_2 but have no predecessors in year t_1 . These “new” clusters represent only the fastest-growing locations among a larger pool of similar areas, many of which did not grow enough to become urban clusters. Including these successful outliers while ignoring the rest would upwardly bias the estimated growth rates for small cities. Our method avoids this by analyzing only components that contain at least one cluster from the start of the period (Step 4).

To produce the final datasets for our analysis, we apply the CCA to the USA and global grids with specific hyper-parameters. We classified urban cells using distinct

²We do so to avoid “new cluster bias” (see below)

criteria for the USA and the global grids. For the USA grids, we use a simple population threshold, classifying a cell as “urban” if it has more than 50 people. For the global grids, we use degree-of-urbanization estimates from the smod grid, classifying a cell as urban if it belongs to a “semi-dense urban cluster” or higher (cell value ≥ 22). Further, we applied a common city threshold to both datasets, filtering out clusters with fewer than 5000 inhabitants. Finally, we excluded from the global cities dataset: (i) Nepal and Myanmar due to data quality issues; (ii) countries with fewer than 50 cities; (iii) countries not present in the urbanization data ([54]). In SI 3.1 and 3.2, we show that our results are robust to reasonable choices of these hyper-parameters.

4.3 Analysis

4.3.1 Theoretical model

We use a simple theoretical model to guide our analysis. This model rests on two simplifying assumptions. First, a city’s size $S(t)$ at time t is a power-law function of its (descending) rank $R(t)$, meaning that the rank-size curve ($x=\log\text{-rank}; y=\log\text{-size}$) is a straight line with slope a_t ³:

$$S(t) \propto R(t)^{-a_t} \quad (2)$$

Second, a city’s growth rate between t and $t + 10$ is a power-law function of its size $S(t)$, meaning that the size-growth curve ($x=\log\text{-size}; y=\log\text{-growth}$) is a straight line with slope b_t :

$$g(t, S(t)) = \frac{S(t+10)}{S(t)} \propto S(t)^{b_t} . \quad (3)$$

Under these assumptions, a_t and b_t are related by a simple equation:

$$a_{t+10} = a_t \cdot (1 + b_t) . \quad (4)$$

For $t_1 > t_0$, this equation generalizes to:

$$a_{t_1} = \exp \left(\log(a_{t_0}) + \sum_{s \in t_0, t_0+10, \dots, t_1-10} \log(1 + b_s) \right) . \quad (5)$$

Furthermore, a_t is linked to the share of the *urban* population living in 1M+ cities, m_t . In fact, the probability of observing a city larger than size x in year t is given by $P(S(t) > x) \propto x^{-1/a_t}$, so:

$$m_t = \int_z^{x_{t,\max}} P(S(t) > x) dx / \int_{x_{t,\min}}^{x_{t,\max}} P(S(t) > x) dx \quad (6)$$

$$= \frac{x_{t,\max}^{1-1/a_t} - z^{1-1/a_t}}{x_{t,\max}^{1-1/a_t} - x_{t,\min}^{1-1/a_t}} . \quad (7)$$

³Note that with this definition the complementary cumulative density function (CCDF) of the city size distribution is $P(S > x) \propto x^{-1/a_t}$. Other authors define the exponent with its reciprocal $\zeta_t = 1/a_t$ so that the CCDF is $P(S > x) \propto x^{-\zeta_t}$.

Here, $x_{t,\max}$ and $x_{t,\min}$ are upper and lower bounds on the size of a country's cities, and $z = 10^6 = 1\text{M}$.

In sum, this theoretical model provides simple closed-form equations relating the growth advantage of large cities (b) to the concentration of population within them (a, m).

4.3.2 Empirical measurement

The above model highlights a and b as the core parameters governing the evolution of urban systems. We estimate these parameters as follows:

- a* The rank-size slope α is our empirical estimate for the parameter a . To obtain α we first estimate the rank-size curve h by fitting a penalized cubic B-spline (with penalty $\lambda = 1$) to city log-rank vs. city log-size data points:

$$\log_{10}(S(t)) = h(\log_{10}(R(t))) . \quad (8)$$

Then we take the absolute value of the mean derivative of h over the log-rank spectrum:

$$\alpha = \frac{1}{r_{\max} - r_{\min}} \left| \int_{r_{\min}}^{r_{\max}} h'(r) dr \right| = \left| \frac{h(r_{\max}) - h(r_{\min})}{r_{\max} - r_{\min}} \right| , \quad r = \log_{10}(R) . \quad (9)$$

- b* The size-growth slope β is our empirical estimate for the parameter b . As above, we obtain β by first estimating the size-growth curve f , a penalized cubic B-spline (with penalty $\lambda = 100$) fit to the city log-size vs. city log-growth data points:

$$\log_{10}(g(t, S(t))) = f(\log_{10}(S(t))) . \quad (10)$$

Then we take the mean derivative of f over the observed log-size spectrum:

$$\beta = \frac{1}{s_{\max} - s_{\min}} \int_{s_{\min}}^{s_{\max}} f'(s) ds = \frac{f(s_{\max}) - f(s_{\min})}{s_{\max} - s_{\min}} , \quad s = \log_{10}(S) . \quad (11)$$

We also estimate size-growth curves for groups of countries (such as Europe or less urbanized countries in 1975). The procedure to estimate a group-level size-growth curve f_g features three steps. First, we normalize each city's growth rate $g(t, S_{ic}(t)) = S_{ic}(t+10)/S_{ic}(t)$ by the average city growth rate of its country c :

$$g_c(t) = \sum_{i \in c} S_{ic}(t+10) / \sum_{i \in c} S_{ic}(t) = \sum_i \frac{S_{ic}(t)}{\sum_{i \in c} S_{ic}(t)} \cdot g(t, S_{ic}(t)) . \quad (12)$$

Second, we pool the normalized growth rates from all countries in the group and fit a single penalized cubic B-spline f_g (with $\lambda = 100$):

$$\log_{10} \left(\frac{g(t, S_{ic}(t))}{g_c(t)} \right) = \bar{f}_g(\log_{10}(S_{ic}(t))) . \quad (13)$$

Independent \ Dependent	Size-growth slope β	
Urban population share	-0.049*** (0.004)	-0.049*** (0.012)
Country fixed effect	No	Yes
Observations	891	891
R^2	0.168	0.545

*p<0.1; **p<0.05; ***p<0.01

Table 2 Association between country urbanization and β . Column (1): pooled specification. Column (2): country fixed effects. The sample is the global cities dataset (Table 1).

Third, we recenter the function $\bar{f}_g$ by the group’s average growth rate:

$$f_g = \bar{f}_g + \frac{1}{|g|} \sum_{c \in g} \log_{10}(g_c(t)) . \quad (14)$$

Other authors use ordinary least squares (OLS) regression to estimate the parameters a and b . Our approach has two advantages over the OLS approach. First, the resulting parameter estimates are more robust to arbitrary design choices, such as the lower threshold for city size. Second, the estimates align more closely with the equations derived from the theoretical model, even when the data deviate from the model’s ideal assumptions (SI Figs. 4 and 5). These advantages can be explained by how each approach calculates average slopes. Our approach calculates the average slope of a given curve by putting equal weights on each part of the x-range. The OLS approach, in contrast, weights parts of the x-range by the density of data that are within them. Because small cities dominate city population data, the OLS slope is disproportionately influenced by the small-city end of the curve. This makes the OLS slope susceptible to variation in the city size lower threshold and less consistent with our theoretical equations. In SI 2.1, we compare spline-based and OLS-based parameter estimates in greater detail.

4.4 Projections

The core idea of our projection method is to take the closed-form equations from Methods 4.3.1 and populate them with our spline-based measurements from Methods 4.3.2. The logic is that the theoretical model captures the form of the relationship between parameters, while our spline-based measurements provide the most accurate values for these parameters (see SI 2.2).

We start by projecting β . We consider four scenarios:

- Proportional growth: We set $\beta_{tc} = 0$ for all $t \geq 2020$ and all countries c .
- Increasing returns: We set $\beta_{tc} = 0.03$ for all $t \geq 2020$ and all countries c .
- Current trend extrapolation: We set $\beta_{tc} = \bar{\beta}_c$ for all $t \geq 2020$ and all countries c . Here, $\bar{\beta}_c$ is the average β for country c over 1975-2025 (as mapped in Figure 1C).

- Our model: We regress β on the urban population share u to obtain $\rho = -0.049$ (Table 2). Then, for $t \geq 2020$, we set:

$$\beta_{tc} = \bar{\beta}_c + \rho \cdot (u_{ct} - \bar{u}_c) . \quad (15)$$

Here, u_{ct} is the projected urban population share for country c from [54] (Methods 4.1) and $\bar{u}_c$ is the average urban population share for country c over 1975-2025.

We project α for $t \geq 2030$ by plugging the projected β s into an empirical version of equation (5):

$$\alpha_t = \exp \left(\log(\alpha_{2020}) + \sum_{s \in 2020, \dots, t-10} \log(1 + \beta_s) \right) . \quad (16)$$

We project the share of *total* population living in 1M+ cities using a two-step approach.

1. First, we project the share of *urban* population in 1M+ cities using an empirical version of equation (6):

$$m_t = \frac{x_{t,\max}^{1-1/\alpha_t} - x_{t,\min}^{1-1/\alpha_t}}{x_{t,\max}^{1-1/\alpha_t} - x_{t,\min}^{1-1/\alpha_t}} . \quad (17)$$

The key challenge in evaluating the right-hand side of this equation is to set plausible values for $x_{t,\min}$ and $x_{t,\max}$, the lower and upper bounds on the sizes of a country's cities. For $x_{t,\min}$, we use the city size lower bound used throughout the paper, $x_{t,\min} = 5000$. For $x_{t,\max}$, the choice is less straightforward because there is no clear upper bound for the size of a country's cities. We assume that $x_{t,\max}$ is proportional to the total urban population of a country U_t , i.e., $x_{t,\max} = \omega \cdot U_t$ where ω is a tunable parameter. We then calibrate ω using historical data. For each country, we select $\omega \in (0.1, 2)$ to minimize the average deviation between the shares m_t calculated using equation (17) and the shares $\hat{m}_t$ observed directly in the data. Aggregated at the regional level, the calibrated estimates m_t closely match the observed data $\hat{m}_t$ (mean absolute error $|m_t - \hat{m}_t|$ below 0.01; see SI Fig. 6).

2. Second, we multiply the projected shares m_t by the projected urban population share u_t to obtain the share of *total* population living in 1M+ cities.

4.5 Code and data availability

All data and source code underlying this paper are publicly available.

Code: To enable exact re-execution, we provide a fully automated, end-to-end pipeline that reproduces the entire analysis with a single command. The pipeline is available on GitHub <https://github.com/ethz-coss/global-city-growth> and archived on Zenodo at the time of publication <https://doi.org/10.5281/zenodo.17339403>. It handles ~ 300 GB of data spanning heterogeneous formats like raster, vector/shapefile, CSV, and Parquet. It executes 100+ tasks in a directed acyclic graph and produces 200+ output tables persisted in PostgreSQL (primary store) and processed

with DuckDB (for fast processing of large tables). It is containerized with Docker. On a MacBook Pro (Apple M2, 32 GB RAM), a full run completes in ~ 5 hours.

Source data: We deposited all redistributable source data at <https://doi.org/10.5281/zenodo.17343655>. Some datasets (IPUMS and NHGIS) are public but not redistributable. The pipeline retrieves them via the IPUMS API. Running the pipeline requires a personal IPUMS API key (see <https://developer.ipums.org/docs/v2/apiprogram/>).

Output data: A full database dump of all aggregate tables resulting from a pipeline run is available at <https://doi.org/10.5281/zenodo.17338968>. The key output datasets — city boundaries and populations for the USA (1850-2020) and the world (1975-2025) + projections of population shares in 1M+ cities by country — are available at <https://doi.org/10.5281/zenodo.17315338>.

4.6 Acknowledgments

Andrea Musso and Dirk Helbing acknowledge support from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation program (833168). Frank Neffke acknowledges financial support from the Austrian Research Promotion Agency (FFG) in the framework of the project ESSENCSE (873927), within the funding program Complexity Science. Diego Rybski acknowledges support from the German Research Foundation (DFG) project UPon (451083179) and project Gropius (511568027). Andrea Musso is grateful to Simone Daniotti, Cesare Carissimo, and Ricardo Hausmann for helpful conversations.

References

- [1] Krugman, P.: Increasing Returns and Economic Geography. *Journal of Political Economy* **99**(3), 483–499 (1991) [2937739](https://doi.org/10.1086/2618625)
- [2] Fujita, M., Krugman, P.R., Venables, A.: *The Spatial Economy: Cities, Regions, and International Trade*. MIT Press, Cambridge, MA (2001)
- [3] Pumain, D., Paulus, F., Vacchiani-Marcuzzo, C., Lobo, J.: An evolutionary theory for interpreting urban scaling laws. *Cybergeo: European Journal of Geography* (2006) <https://doi.org/10.4000/cybergeo.2519>
- [4] Paulus, F.: *Coévolution dans les systèmes de villes : croissance et spécialisation des aires urbaines françaises de 1950 à 2000*. PhD thesis, Université Panthéon-Sorbonne - Paris I (2004)
- [5] Pumain, D.: Une Théorie Géographique Pour La Loi De Zipf. *Region et Developpement* **36**, 31–54 (2012)
- [6] Bettencourt, L.M.A., Lobo, J., Helbing, D., Kühnert, C., West, G.B.: Growth, innovation, scaling, and the pace of life in cities. *Proceedings of the National*

- Academy of Sciences **104**(17), 7301–7306 (2007) <https://doi.org/10.1073/pnas.0610172104>
- [7] Bettencourt, L.M.A., Samaniego, H., Youn, H.: Professional diversity and the productivity of cities. *Scientific Reports* **4**(1), 5393 (2014) <https://doi.org/10.1038/srep05393>
 - [8] Combes, P.-P., Duranton, G., Gobillon, L.: Spatial wage disparities: Sorting matters! *Journal of Urban Economics* **63**(2), 723–742 (2008) <https://doi.org/10.1016/j.jue.2007.04.004>
 - [9] Keuschnigg, M., Mutgan, S., Hedström, P.: Urban scaling and the regional divide. *Science Advances* **5**(1), 0042 (2019) <https://doi.org/10.1126/sciadv.aav0042>
 - [10] Balland, P.-A., Jara-Figueroa, C., Petralia, S.G., Steijn, M.P.A., Rigby, D.L., Hidalgo, C.A.: Complex economic activities concentrate in large cities. *Nature Human Behaviour* **4**(3), 248–254 (2020) <https://doi.org/10.1038/s41562-019-0803-3>
 - [11] Gomez-Lievano, A., Patterson-Lomba, O., Hausmann, R.: Explaining the prevalence, scaling and variance of urban phenomena. *Nature Human Behaviour* **1**(1), 1–6 (2016) <https://doi.org/10.1038/s41562-016-0012>
 - [12] Sutton, J.: Gibrat’s Legacy. *Journal of Economic Literature* **35**(1), 40–59 (1997) [2729692](https://doi.org/10.2307/2729692)
 - [13] Eeckhout, J.: Gibrat’s Law for (All) Cities. *American Economic Review* **94**(5), 1429–1451 (2004) <https://doi.org/10.1257/0002828043052303>
 - [14] Eaton, J., Eckstein, Z.: Cities and growth: Theory and evidence from France and Japan. *Regional Science and Urban Economics* **27**(4-5), 443–474 (1997)
 - [15] González-Val, R., Lanaspá, L., Sanz-Gracia, F.: New Evidence on Gibrat’s Law for Cities. *Urban Studies* **51**(1), 93–115 (2014) <https://doi.org/10.1177/0042098013484528>
 - [16] Henderson, J.V., Wang, H.G.: Urbanization and city growth: The role of institutions. *Regional Science and Urban Economics* **37**(3), 283–313 (2007)
 - [17] Louf, R., Barthélemy, M.: How congestion shapes cities: From mobility patterns to scaling. *Scientific Reports* **4**(1), 5561 (2014) <https://doi.org/10.1038/srep05561>
 - [18] Barthélemy, M.: *The Structure and Dynamics Of Cities*. Cambridge University Press, Cambridge (2016)
 - [19] Oliveira, M., Bastos-Filho, C., Menezes, R.: The scaling of crime concentration in cities. *PLOS ONE* **12**(8), 0183110 (2017) <https://doi.org/10.1371/journal.pone.0183110>

- [20] Bettencourt, L.M.A., Lobo, J., Strumsky, D., West, G.B.: Urban Scaling and Its Deviations: Revealing the Structure of Wealth, Innovation and Crime across Cities. *PLOS ONE* **5**(11), 13541 (2010) <https://doi.org/10.1371/journal.pone.0013541>
- [21] Rocha, L.E.C., Thorson, A.E., Lambiotte, R.: The Non-linear Health Consequences of Living in Larger Cities. *Journal of Urban Health* **92**(5), 785–799 (2015) <https://doi.org/10.1007/s11524-015-9976-x>
- [22] Bilal, U., de Castro, C.P., Alfaro, T., Barrientos-Gutierrez, T., Barreto, M.L., Leveau, C.M., Martinez-Folgar, K., Miranda, J.J., Montes, F., Mullachery, P., Pina, M.F., Rodriguez, D.A., dos Santos, G.F., Andrade, R.F.S., Diez Roux, A.V.: Scaling of mortality in 742 metropolitan areas of the Americas. *Science Advances* **7**(50), 6325 (2021) <https://doi.org/10.1126/sciadv.abl6325>
- [23] Sarkar, S.: Urban scaling and the geographic concentration of inequalities by city size. *Environment and Planning B: Urban Analytics and City Science* **46**(9), 1627–1644 (2019) <https://doi.org/10.1177/2399808318766070>
- [24] Duranton, G., Puga, D.: Chapter 48 - Micro-Foundations of Urban Agglomeration Economies. In: Henderson, J.V., Thisse, J.-F. (eds.) *Handbook of Regional and Urban Economics. Cities and Geography*, vol. 4, pp. 2063–2117. Elsevier, ??? (2004). [https://doi.org/10.1016/S1574-0080\(04\)80005-1](https://doi.org/10.1016/S1574-0080(04)80005-1)
- [25] Schiavina, M., Melchiorri, M., Pesaresi, M., Politis, P., Carneiro Freire, S.M., Maffenini, L., Florio, P., Ehrlich, D., Goch, K., Carioli, A., Uhl, J., Tommasi, P., Kemper, T.: GHSL Data Package 2023. Publications Office of the European Union, Luxembourg (2023). <https://doi.org/10.2760/098587>
- [26] Ruggles, S., Nelson, M.A., Sobek, M., Fitch, C.A., Goeken, R., Hacker, J.D., Roberts, E., Warren, J.R.: IPUMS ancestry full count data: Version 4.0 (2024) <https://doi.org/10.18128/D014.V4.0>
- [27] Manson, S., Schroeder, J., Riper, D.V., Knowles, K., Kugler, T., Roberts, F., Ruggles, S.: IPUMS National Historical Geographic Information System: Version 19.0. IPUMS, Minneapolis, MN (2024). <https://doi.org/10.18128/D050.V19.0>
- [28] Berkes, E., Karger, E., Nencka, P.: The census place project: A method for geolocating unstructured place names. *Explorations in Economic History* **87**, 101477 (2023) <https://doi.org/10.1016/j.eeh.2022.101477>
- [29] Pumain, D., Swerts, E., Cottineau, C., Vacchiani-Marcuzzo, C., Ignazzi, C.A., Bretagnolle, A., Delisle, F., Cura, R., Lizzi, L., Baffi, S.: Multilevel comparison of large urban systems. *Cybergeo: European Journal of Geography* (2015) <https://doi.org/10.4000/cybergeo.26730>
- [30] Guerin-Pace, F.: Rank-Size Distribution and the Process of Urban Growth. *Urban*

Studies **32**(3), 551–562 (1995)

- [31] Desmet, K., Rappaport, J.: The settlement of the United States, 1800–2000: The long transition towards Gibrat’s law. *Journal of Urban Economics* **98**(C), 50–68 (2017)
- [32] Michaels, G., Rauch, F., Redding, S.J.: Urbanization and Structural Transformation *. *The Quarterly Journal of Economics* **127**(2), 535–586 (2012) <https://doi.org/10.1093/qje/qjs003>
- [33] Rozenfeld, H.D., Rybski, D., Andrade, J.S., Batty, M., Stanley, H.E., Makse, H.A.: Laws of population growth. *Proceedings of the National Academy of Sciences* **105**(48), 18702–18707 (2008) <https://doi.org/10.1073/pnas.0807435105>
- [34] Black, D., Henderson, V.: Urban evolution in the USA. *Journal of Economic Geography* **3**(4), 343–372 (2003) <https://doi.org/10.1093/jeg/lbg017>
- [35] Soo, K.T.: Zipf’s Law and Urban Growth in Malaysia. *Urban Studies* **44**(1), 1–14 (2007) [43084333](https://doi.org/10.1080/00439860600584333)
- [36] Cottineau, C.: MetaZipf. A dynamic meta-analysis of city size distributions. *PLOS ONE* **12**(8), 0183919 (2017) <https://doi.org/10.1371/journal.pone.0183919>
- [37] Verbavatz, V., Barthelemy, M.: The growth equation of cities. *Nature* **587**(7834), 397–401 (2020) <https://doi.org/10.1038/s41586-020-2900-x>
- [38] Zipf, G.K.: Human Behavior and the Principle of Least Effort. Human Behavior and the Principle of Least Effort, p. 573. Addison-Wesley Press, Oxford, England (1949)
- [39] Batty, M.: Rank clocks. *Nature* **444**(7119), 592–596 (2006) <https://doi.org/10.1038/nature05302>
- [40] Rybski, D., Ciccone, A.: Auerbach, Lotka, and Zipf: Pioneers of power-law city-size distributions. *Archive for History of Exact Sciences* **77**(6), 601–613 (2023) <https://doi.org/10.1007/s00407-023-00314-0>
- [41] Saichev, A., Malevergne, Y., Sornette, D.: Theory of Zipf’s Law And Beyond. *Lecture Notes in Economics and Mathematical Systems*, vol. 632. Springer, Berlin, Heidelberg (2010). <https://doi.org/10.1007/978-3-642-02946-2>
- [42] Makse, H.A., Havlin, S., Stanley, H.E.: Modelling urban growth patterns. *Nature* **377**(6550), 608–612 (1995) <https://doi.org/10.1038/377608a0>
- [43] Arcaute, E., Hatna, E., Ferguson, P., Youn, H., Johansson, A., Batty, M.: Constructing cities, deconstructing scaling laws. *Journal of The Royal Society Interface* **12**(102), 20140745 (2015) <https://doi.org/10.1098/rsif.2014.0745>

- [44] Láng-Ritter, J., Keskinen, M., Tenkanen, H.: Global gridded population datasets systematically underrepresent rural population. *Nature Communications* **16**(1), 2170 (2025) <https://doi.org/10.1038/s41467-025-56906-7>
- [45] Reia, S.M., Rao, P.S.C., Barthelemy, M., Ukkusuri, S.V.: Spatial structure of city population growth. *Nature Communications* **13**(1), 5931 (2022) <https://doi.org/10.1038/s41467-022-33527-y>
- [46] Pumain, D., Moriconi-Ebrard, F.: City size distributions and metropolisation. *GeoJournal* **43**(4), 307–314 (1997) <https://doi.org/10.1023/A:1006833025206>
- [47] Ribeiro, H.V., Oehlers, M., Moreno-Monroy, A.I., Kropp, J.P., Rybski, D.: Association between population distribution and urban GDP scaling. *PLOS ONE* **16**(1), 0245771 (2021) <https://doi.org/10.1371/journal.pone.0245771>
- [48] Gomez-Lievano, A., Youn, H., Bettencourt, L.M.A.: The Statistics of Urban Scaling and Their Connection to Zipf’s Law. *PLOS ONE* **7**(7), 40393 (2012) <https://doi.org/10.1371/journal.pone.0040393>
- [49] United Nations, Department of Economic and Social Affairs, Population Division: World Population Prospects 2024. United Nations (2024). <https://doi.org/10.18356/9789211065138>
- [50] Klein Goldewijk, K.: History Database of the Global Environment (HYDE) 3.3. Utrecht University / PBL Netherlands Environmental Assessment Agency (2023). <https://doi.org/10.24416/UU01-AEZZIT>
- [51] Gapminder: Systema Globalis (2023)
- [52] Gapminder: Population V7 (2022)
- [53] PBL Netherlands Environmental Assessment Agency, Gapminder, United Nations, Department of Economic and Social Affairs, Population Division, Our World in Data: Population — HYDE, Gapminder, UN — Long-Run Data. Dataset (2024)
- [54] Chen, S., Huang, Q., Muttarak, R., Fang, J., Liu, T., He, C., Liu, Z., Zhu, L.: Updating global urbanization projections under the Shared Socioeconomic Pathways. *Scientific Data* **9**(1), 137 (2022) <https://doi.org/10.1038/s41597-022-01209-5>
- [55] Schvitz, G., Rüegger, S., Girardin, L., Cederman, L.-E., Weidmann, N., Gleditsch, K.S.: Mapping the international system, 1886-2017: The cshapes 2.0 dataset. *Journal of Conflict Resolution* **66**(1), 144–161 (2022) <https://doi.org/10.1177/00220027211013563>
- [56] Batista e Silva, F., Freire, S., Schiavina, M., Rosina, K., Marín-Herrera, M.A.,

- Ziemba, L., Craglia, M., Koomen, E., Lavalle, C.: Uncovering temporal changes in Europe's population density patterns using a data fusion approach. *Nature Communications* **11**(1), 4631 (2020) <https://doi.org/10.1038/s41467-020-18344-5>
- [57] Hoshen, J., Kopelman, R.: Percolation and cluster distribution. I. Cluster multiple labeling technique and critical concentration algorithm. *Physical Review B* **14**(8), 3438–3445 (1976) <https://doi.org/10.1103/PhysRevB.14.3438>