Laws of population growth

Hernán D. Rozenfelda, Diego Rybski, José S. Andrade, Jr.b, Michael Batty, H. Eugene Stanley, and Hernán A. Makseab,1

aLevich Institute and Physics Department, City College of New York, New York, NY 10031; bDepartamento de Física, Universidade Federal do Ceará, 60451-970 Fortaleza, Ceará, Brazil; cCentre for Advanced Spatial Analysis, University College London, 1-19 Torrington Place, London WC1E 6BT, United Kingdom; and dCenter for Polymer Studies and Physics Department, Boston University, Boston, MA 02215

Edited by Michael F. Goodchild, University of California, Santa Barbara, CA, and approved September 29, 2008 (received for review July 31, 2008)

An important issue in the study of cities is defining a metropolitan area, because different definitions affect conclusions regarding the statistical distribution of urban activity. A commonly employed method of defining a metropolitan area is the Metropolitan Statistical Areas (MSAs), based on rules attempting to capture the notion of city as a functional economic region, and it is performed by using experience. The construction of MSAs is a time-consuming process and is typically done only for a subset (a few hundreds) of the most highly populated cities. Here, we introduce a method to designate metropolitan areas, denoted “City Clustering Algorithm” (CCA). The CCA is based on spatial distributions of the population at a fine geographic scale, defining a city beyond the scope of its administrative boundaries. We use the CCA to examine Gibrat’s law of proportional growth, which postulates that the mean and standard deviation of the growth rate of cities are constant, independent of city size. We find that the mean and standard deviation of a cluster by utilizing the CCA exhibits deviations from Gibrat’s law, and that the standard deviation decreases as a power law with respect to the city size. The CCA allows for the study of the underlying process leading to these deviations, which are shown to arise from the existence of long-range spatial correlations in population growth. These results have sociopolitical implications, for example, for the location of new economic development in cities of varied size.

Results

In recent years there has been considerable work on how to define cities and how the different definitions affect the statistical distribution of urban activity (1, 2). This is a long-standing problem in spatial analysis of aggregated data sources, referred to as the “modifiable areal unit problem” or the “ecological fallacy” (3, 4), where different definitions of spatial units based on administrative or governmental boundaries give rise to inconsistent conclusions with respect to explanations and interpretations of data at different scales. The conventional method of defining human agglomerations is through the Metropolitan Statistical Areas (MSAs) (1, 2, 5–7), which are subject to socioeconomic factors. The MSA has been of indubitable importance for the analysis of population growth, and is constructed manually case-by-case based on subjective judgment (MSAs are defined by starting from a highly populated central area and adding its surrounding counties if they have social or economical ties).

In this report, we propose a way to measure the extent of human agglomerations based on clustering techniques by using a fine geographical grid, covering both urban and rural areas. In this view, “cities” represent clusters of population, i.e., adjacent populated geographical spaces. Our algorithm, the “city clustering algorithm” (CCA), allows for an automated and systematic way of building population clusters based on the geographical location of people. The CCA has one parameter (the cell size) that is useful for the study of human agglomerations at different length scales, similar to the level of aggregation in the context of social sciences. We show that the CCA allows for the study of the origin of statistical properties of population growth. We use the CCA to analyze the postulates of Gibrat’s law of proportional growth applied to cities, which assumes that the mean and standard deviation of the growth rates of cities are constant. We show that population growth at a fine geographical scale for different urban and regional systems at country and continental levels (Great Britain, the United States, and Africa) deviates from Gibrat’s law. We find that the mean and standard deviation of population growth rates decrease with population size, in some cases following a power-law behavior. We argue that the underlying demographic process leading to the deviations from Gibrat’s law can be modeled from the existence of long-range spatial correlations in the growth of the population, which may arise from the concept that “development attracts further development.” These results have implications for social policies, such as those pertaining to the location of new economic development in cities of different sizes. The present results imply that, on average, the greatest growth rate occurs in the smallest places where there is the greatest risk of failure (larger fluctuations). A corollary is that the safest growth occurs in the largest places having less likelihood for rapid growth.

The analyzed data consist of the number of inhabitants, \( n(t) \), in each cell \( i \) of a fine geographical grid at a given time, \( t \). The cell size varies for each dataset used in this study. We consider three different geographic scales: on the smallest scale, the area of study is Great Britain (GB: England, Scotland and Wales), a highly urbanized country with a population of 58.7 million in 2007, and an area of 0.23 million km\(^2\). The grid is composed of 5.75 million cells of 2 km by 2 km. Therefore, the analyzed datasets of Great Britain and the United States are populated-places datasets, with population counts defined at points in a grid. Because there could be some distortions in the true residential population involved at the finest grid resolution, we perform our analysis by investigating the statistical properties as a function of the grid size by coarse-graining the data as explained in Information on the Datasets. At the largest scale, we analyze the continent of Africa, composed of 53 countries with a total population of 933 million in 2007, and an area of 23 million km\(^2\). The original USA data consists of 59,456 sites defined by Federal Information Processing Standards (FIPS) associated with a corresponding population provided by the U.S. Census Bureau (9), which is then coarse-grained to a grid of 2 km by 2 km. Therefore, the analyzed datasets of Great Britain and the United States are populated-places datasets, with population counts defined at points in a grid. Because there could be some distortions in the true residential population involved at the finest grid resolution, we perform our analysis by investigating the statistical properties as a function of the grid size by coarse-graining the data as explained in Information on the Datasets. At the largest scale, we analyze the continent of Africa, composed of 53 countries with a total population of 933 million in 2007, and an area of 30.34 million km\(^2\). These data are gridded with less resolution by 0.50 million cells of approximately 7.74 km by 7.74 km (10). More detailed information about these datasets is found in Information on the Datasets (all the datasets studied in this article are available at http://lev.ccny.cuny.edu/~hmakse/cities/city_data.zip).

Results

Fig. 1 illustrates operation of the CCA. To identify urban clusters, we require connected cells to have nonzero population. We...
start by selecting an arbitrary populated cell (final results are independent of the choice of the initial cell). Iteratively, we then grow a cluster by adding nearest neighbors of the boundary cells with a population strictly $> 0$, until all neighbors of the boundary are unpopulated. We repeat this process until all populated cells have been assigned to a cluster. This technique was introduced to model forest fire dynamics (11) and is termed the “burning algorithm.” Because one can think of each populated cell as a burning tree.

The population $S_i(t)$ of cluster $i$ at time $t$ is the sum of the populations $n_j^{(t)}(t)$ of each cell $j$ within it, $S_i(t) = \sum_{j=1}^{N_i} n_j^{(t)}(t)$, where $N_i$ is the number of cells in the cluster. Results of the CCA are shown in Fig. 1B, representing the urban cluster surrounding the City of London (red cluster overlaying a satellite image, see http://lev.ccny.cuny.edu/~hmakse/cities/london.gif for an animated image of Fig. 1B). Fig. 1C depicts all the clusters of Great Britain, indicating the large variability in their population and size.

The CCA allows the analysis of the population clusters at different length scales by coarse-graining the grid and applying the CCA to the coarse-grained dataset (see Information on the Datasets for details on coarse-graining the data). At larger scales, disconnected areas around the edge of a cluster could be added into the cluster. This is justified when, for example, a town is divided by a wide highway or a river.

Tables S1 and S2 in supporting information (SI) Appendix show a detailed comparison between the urban clusters obtained with the CCA applied to the United States in 1990, and the results obtained from the analysis of MSAs from the US Census Bureau used in previous studies of population growth (5–7). We observe that the MSAs considered in ref. 5 are similar to the clusters obtained with the CCA with a cell size of 4 km by 4 km or 8 km by 8 km. In particular, the population sizes of the clusters have the same order of magnitude as the MSAs. However, for large cities the MSAs from the data of ref. 6 seem to be mostly comparable to our results for cell sizes of 2 km by 2 km or 4 km by 4 km.

Use of the CCA permits a systematic study of cluster dynamics. For instance, clusters may expand or contract, merge or split between two considered times, as illustrated in Fig. 2. We quantify these processes by measuring the probability distribution of the temporal changes in the clusters for the data of Great Britain. We find that when the cell size is 2.2 km by 2.2 km, 84% of the clusters evolve from 1981 to 1991 following the first 3 cases presented in Fig. 2 (no change, expansion, or reduction), 6% of the clusters merge from 2 clusters into one in 1991, and 3% of the clusters split into 2 clusters.

Next, we apply the CCA to study the dynamics of population growth by investigating Gibrat’s law, which postulates that the mean and standard deviation of growth rates are constant (1, 2, 5, 7, 12). The conventional method (1, 2, 7) is to assume that
the populations of a given city or cluster \( i \), at times \( t_0 \) and \( t_1 > t_0 \), are related by

\[
S_1 = R(S_0)S_0,
\]

where \( S_0 \equiv S(t_0) = \sum_i n_i(t_0) \) and \( S_1 \equiv S(t_1) = \sum_i n_i(t_1) \) are the initial and final populations of cluster \( i \), respectively, and \( R(S_0) \) is the positive growth factor, which varies from cluster to cluster. Following the literature in population dynamics (1, 2, 5, 7), we define the population growth rate of a cluster as

\[
r(S_0) = \ln(S_1)/t - \ln(S_0)/t_0 = \ln(S_1)/t_1 - \ln(S_0)/t_0.
\]

The CCA allows for a study of the growth rates as a function of the scale of observation, by changing the size of the grid. We find (SI Appendix. Section II) that the data for GB are approximately invariant under coarse-graining the grid at different levels for both the mean and standard deviation. When the data of the United States are aggregated spatially from cell size 2 km to 8 km, the scaling of the mean rates crosses over to a flat behavior closer to Gibrat’s law. At the scale of 8 km the mean is approximately constant (with fluctuations). However, we find that, at this scale, all cities in the northeastern the United States spanning from Boston to Washington, DC, form a single cluster. Despite these differences, the scaling of the standard deviation for the United States holds approximately invariant even up to the large scale of observation of 8 km.

Fig. 3 shows the dependence of the standard deviation \( \sigma(S_0) \) on the initial population \( S_0 \). On average, fluctuations in the growth rate of large cities are smaller than for small cities in contrast to Gibrat’s law. This result can be approximated over many orders of magnitude by the power law,

\[
\sigma(S_0) \sim S_0^{-\beta},
\]

where \( \beta \) is the standard deviation exponent. We carry out an OLS regression analysis and find that \( \beta_{USA} = 0.20 \pm 0.06 \). The presence of a power law implies that fluctuations in the growth process are statistically self-similar at different scales, for populations ranging from \( \sim 1,000 \) to \( \sim 10 \) million according to Fig. 3B.

Fig. 4 shows the analysis of the growth rate of the population clusters of Great Britain from gridded databases (8) with a cell size of 2 km by 2 km at \( t_0 = 1981 \) and \( t_1 = 1991 \). The average growth rate depicted in Fig. 4A comprises large fluctuations as a function of \( S_0 \), especially for smaller populations. However, a slight decrease with population seems evident from rates around \( \langle r \rangle \approx 0.008 \pm 0.001 \) with \( S_0 \approx 10^5 \) dropping to zero or even negative values for the largest populations,

\[
S_0 \approx 10^6.\]

We find that 3,556 clusters with population at approximately \( S_0 \approx 10^4 \) exhibit negative growth rates as well. Thus, the mean rates are plotted on a semilogarithmic scale in Fig. 4A. When considering intermediate populations ranging from \( S_0 = 3,000 \) to \( S_0 = 3 \times 10^5 \), the data seem to be following approximately a power law with \( \alpha_{GB} = 0.17 \pm 0.05 \) from OLS regression analysis, as shown in Fig. 4A Inset. Fig. 4B shows the standard deviation for GB, \( \sigma(S_0) \), exhibiting deviations from Gibrat’s law having a tendency to decrease with population according to Eq. 3 and a standard deviation exponent, \( \beta_{GB} = 0.27 \pm 0.04 \), obtained with OLS technique.
Next, we analyze the population growth in Africa during the period from 1960 to 1990 (10). In this case, the population data are based on a larger cell size, so we evaluate the data cell by cell (without the application of the CCA). Despite the differences in the economic and urban development of Africa, Great Britain, and the USA, we find that the mean and standard deviation of the growth rate in Africa display similar scaling as found for the United States and Great Britain. In Fig. 5A we show the results for the growth rate in Africa when the grid is coarse-grained with a cell size of 77 km by 77 km. We find a decrease of the growth rate from \( \langle r(S_0) \rangle \approx 0.1 \) to \( \langle r(S_0) \rangle \approx 0.01 \) between populations \( S_0 \approx 10^3 \) and \( S_0 \approx 10^6 \), respectively. All populations have positive growth rates. A log-log plot of the mean rates shown in Fig. 5A reveals a power-law scaling \( \langle r(S_0) \rangle \sim S_0^{-\alpha_{Af}} \), with \( \alpha_{Af} = 0.21 \pm 0.05 \) from OLS regression analysis. The standard deviation (Fig. 5B) satisfies Eq. 3 with a standard deviation exponent \( \beta_{Af} = 0.19 \pm 0.04 \).

The CCA allows for a study of the origin of the observed behavior of the growth rates by examining the dynamics and spatial correlations of the population of cells. To this end, we first generate a surrogate dataset that consists of shuffling two randomly chosen populated cells, \( n_{ij}^{(d)}(t_0) \) and \( n_{ik}^{(d)}(t_0) \), at time \( t_0 \). This swapping process preserves the probability distribution of \( n_{ij}^{(d)} \) but destroys any spatial correlations among the population cells. Fig. 4C shows the results of the randomization of the Great Britain dataset, indicating power-law scaling in the tail of \( \sigma(S_0) \) with standard deviation exponent \( \beta_{rand} = 1/2 \). This result can be interpreted in terms of the uncorrelated nature of the randomized dataset (SI Appendix Section III). We consider that the population of each cell \( j \) increases by a random amount \( \delta_j \) with mean value \( \bar{\delta} \) and variance \( \langle (\delta - \bar{\delta})^2 \rangle = \Delta^2 \), and that \( r \ll 1 \), then \( n_{ij}^{(d)}(t_1) = n_{ij}^{(d)}(t_0) + \delta_j \). Therefore, the population of a cluster at time \( t_1 \) can be written as

\[
S_1 = S_0 + \sum_{j=1}^{N_j} \delta_j. \tag{4}
\]

It can be shown that (SI Appendix Section III):

\[
\langle S_{1j}^2 \rangle = \langle S_{0j}^2 \rangle + \sum_{j=1}^{N_j} \sum_{k=1}^{N_j} \langle (\delta_j - \bar{\delta})(\delta_k - \bar{\delta}) \rangle. \tag{5}
\]

Randomly shuffling population cells destroys the correlations, leading to \( \langle (\delta_j - \bar{\delta})(\delta_k - \bar{\delta}) \rangle = \Delta^2 \delta_{jk} \) (where \( \delta_{jk} \) is the Kronecker delta function) which implies \( \beta_{rand} = 1/2 \) (16) (see SI Appendix Section III).

The fact that \( \beta \) lies below the random exponent \( \beta_{rand} = 1/2 \) for all the analyzed data suggests that the dynamics of the population cells display spatial correlations, which are eliminated in
correlation exponent are asymptotically of a scale-invariant form characterized by a growth rate of clusters in Africa versus the initial size of population \( S_0 \). The straight dashed line shows a power-law fit with exponent \( \gamma = 0.21 \pm 0.05 \), obtained by using OLS regression. (B) Standard deviation of the growth rate in Africa. The straight line corresponds to power-law fit by using OLS providing the exponent \( \beta_{\text{fit}} = 0.19 \pm 0.04 \).

The cells are not occupied randomly but spatial correlations arise, because when the population in one cell increases, the probability of growth in an adjacent cell also increases. That is, development attracts further development, an idea that has been used to model the spatial distribution of urban patterns (17). Indeed these ideas are related to the study of the origin of power laws in complex systems (18, 19).

When we analyze the populated cells, we indeed find that spatial correlations in the incremental population of the cells, \( \delta_j \), are asymptotically of a scale-invariant form characterized by a correlation exponent \( \gamma \),

\[
\langle (\delta_j - \bar{\delta}) (\delta_k - \bar{\delta}) \rangle \sim \frac{\delta^2}{|\bar{x}_j - \bar{x}_k|^\gamma},
\]

where \( \bar{x}_j \) is the location of cell \( j \). For Great Britain we find \( \gamma = 0.93 \pm 0.08 \) (see Fig. 4D). In SI Appendix: Section III, we show that power-law correlations in the fluctuations at the cell level, Eq. 6, lead to a standard deviation exponent \( \beta = \gamma/4 \). For \( \gamma = 2 \), the dimension of the substrate, we recover \( \beta_{\text{rand}} = 1/2 \) (larger values of \( \gamma \) result in the same \( \beta \) because when \( \gamma > 2 \) correlations become irrelevant). If \( \gamma = 0 \), the standard deviation of the population growth rates has no dependence on the population size (\( \beta = 0 \)), as stated by Gibrat’s law, stating that the standard deviation does not depend in the cluster size. In the case of Great Britain, \( \gamma = 0.93 \pm 0.08 \) gives \( \beta = 0.23 \pm 0.02 \) approximately consistent with the measured value \( \beta_{\text{GB}} = 0.27 \pm 0.04 \), within the error bars. This observation suggests that the underlying demographic process leading to the scaling in the standard deviation can be modeled as arising from the long-range correlated growth of population cells.

Discussion

Our results suggest the existence of scale-invariant growth mechanisms acting at different geographical scales. Furthermore, Eq. 3 is similar to what is found for the growth of firms and other macroeconomic indicators (16, 20). Thus, our results support the existence of an underlying link between the fluctuation dynamics of population growth and various economic indicators, implying considerable unevenness in economic development in different population sizes. City growth is driven by many processes of which population growth and migration is only one. Our study captures only the growth of population, but not economic growth per se.

Many cities grow economically while losing population and, thus, the processes we imply are those that influence a changing population. Our assumption is that population change is an indicator of city growth or decline and, therefore, we have based our studies on population-clustering techniques. Alternatively, the MSAs provides a set of rules that try to capture the idea of city as a functional economic region.

The results we obtain show scale-invariant properties that we have modeled by using long-range spatial correlations between the population of cells. That is, strong development in an area attracts more development in its neighborhood and much beyond. A key finding is that small places exhibit larger fluctuations than large places. The implications for locating activity in different places are that there is a greater probability of larger growth in small places, but also a greater probability of larger decline. Opportunity must be weighed against the risk of failure.

One may take these ideas to a higher level of abstraction to study cell-to-cell flows (migration, commuting, etc.) gridded at different levels. As a consequence one may define population clusters, or MSAs, in terms of functional linkages between neighboring cells. In addition one may relax some conditions imposed in the CCA. Here, we consider a cell to be part of a cluster only if its population is strictly >0. In SI Appendix: Section V, we relax this condition and study the robustness of the CCA when cells of a higher population than 0 (for instance, 5 and 20) are allowed into clusters and find that, even though small clusters present a slight deviation, the overall behavior of the growth rate and standard deviation is conserved.

Materials and Methods

Information on the Datasets. The datasets analyzed in this article were obtained from the web sites http://census.ac.uk; http://www.esri.com/; and http://na.unep.net/datasets/datalist.php, for Great Britain, the United States, and Africa, respectively, and can be downloaded from http://lev.ccny.cuny.edu/~hnakse/cities/city_data.zip.

The datasets consist of a list of populations at specific coordinates at 2 time steps, \( s_0 \) and \( t_1 \). A graphical representation of the data can be seen in Fig. 1C for Great Britain where each point represents a data point directly extracted from the dataset.

To perform the CCA at different scales we coarse-grain the datasets. For this purpose, we overlay a grid on the corresponding map (United States, Great Britain, or Africa) with the desired cell size (e.g., 2 km by 2 km or 4 km by 4 km for the United States). Then, the population of each cell is calculated as the sum of the populations of points (obtained from the original dataset) that fall into this cell.

Table 1 shows information on the datasets and results on United States, Great Britain, and Africa for the cell size used in the main text as well as some of the exponents obtained in our analysis.

Calculation of \( r(S_0) \) and \( \sigma(S_0) \) and Methodology

The average growth rate, \( \langle r(S_0) \rangle = \ln(S_1/S_0) \), and the standard deviation, \( \sigma(S_0) = \sqrt{\langle (r(S_0))^2 \rangle - \langle r(S_0) \rangle^2} \), are defined as follows. If we call \( F(r(S_0)) \) the conditional probability distribution of finding a cluster with growth rate \( r(S_0) \) with the condition of initial population \( S_0 \), then we can obtain \( r(S_0) \) and \( \sigma(S_0) \) through,

\[
\langle r(S_0) \rangle = \int r F(r(S_0)) dr,
\]

[7]
Table 1. Characteristics of datasets and summary of results

<table>
<thead>
<tr>
<th>Data</th>
<th>No. of cells</th>
<th>t₀</th>
<th>t₁</th>
<th>Average growth rate, %</th>
<th>Cell size</th>
<th>No. of clusters</th>
<th>α</th>
<th>β</th>
</tr>
</thead>
<tbody>
<tr>
<td>USA</td>
<td>1.86 mill</td>
<td>1990</td>
<td>2000</td>
<td>0.9</td>
<td>2 km by 2 km</td>
<td>30,210</td>
<td>0.28 ± 0.08</td>
<td>0.20 ± 0.06</td>
</tr>
<tr>
<td>GB</td>
<td>0.10 mill</td>
<td>1981</td>
<td>1991</td>
<td>0.3</td>
<td>2.2 km by 2.2 km</td>
<td>10,178</td>
<td>0.17 ± 0.05</td>
<td>0.27 ± 0.04</td>
</tr>
<tr>
<td>Africa</td>
<td>2,216</td>
<td>1960</td>
<td>1990</td>
<td>4</td>
<td>77 km by 77 km</td>
<td>3,988</td>
<td>0.21 ± 0.05</td>
<td>0.19 ± 0.04</td>
</tr>
</tbody>
</table>

and

\[ r(S_i(t)) = \int r(S_i) dS. \]  

Once \( r(S_i) \) and \( r(S) \) are calculated for each cluster, we perform a nonparametric regression analysis (13, 14), a technique broadly used in the literature of population dynamics. The idea is to provide an estimate for the relationship between the growth rate and \( S_i \) and between the standard deviation and \( S_i \). Following the methods explained in ref. 14, we apply the Nadaraya–Watson method to calculate an estimate for the growth rate, \( \Sigma_i k_i (S_i - S_i(t_0)) r(S_i) / \Sigma_i k_i (S_i - S_i(t_0)) \), where

\[ k_i (S_i - S_i(t_0)) = \frac{\exp(-h^2 (\ln S_i - \ln S_i(t_0))^2)}{\sqrt{2\pi h}}, \quad h = 0.5 \]  

Finally, we compute the 95% confidence bands (calculated from 500 random samples with replacement) to estimate the amount of statistical error in our results (13). The bootstrapping technique was applied by sampling as many data points as the number of clusters and performing the nonparametric regression on the sampled data. By performing 500 realizations of the bootstrapping algorithm and extracting the so-called \( \alpha/2 \) (\( \alpha \) is not related to the growth rate exponent) quantiles we obtain the 95% confidence bands.

To obtain the exponents \( \alpha \) and \( \beta \) of the power-law scalings for \( r(S_i) \) and \( \sigma(S_i) \), respectively, we perform an OLS regression analysis (15). More specifically, to obtain the exponent \( \beta \) from Eq. 3, we first linearize the data by considering the logarithm of the independent and dependent variables so that Eq. 3 becomes \( \ln r(S_i) \sim \beta \ln S_i \). Then, we apply a linear OLS regression that leads to the exponent

\[
\beta = \frac{N_i \sum_i^{N_i} \ln S_i(t_0) \ln r(S_i(t_0)) - \sum_i^{N_i} \ln S_i(t_0) \sum_i^{N_i} \ln r(S_i(t_0))}{N_i \sum_i^{N_i} (\ln S_i(t_0))^2 - (\sum_i^{N_i} \ln S_i(t_0))^2},
\]

where \( N_i \) is the number of clusters found by using the CCA. Analogously, we obtain the exponent \( \alpha \) by linearizing \( |r(S_i)| \) and calculating

\[
\alpha = \frac{N_i \sum_i^{N_i} \ln S_i(t_0) |r(S_i(t_0))| - \sum_i^{N_i} \ln S_i(t_0) \sum_i^{N_i} \ln |r(S_i(t_0))|}{N_i \sum_i^{N_i} (\ln S_i(t_0))^2 - (\sum_i^{N_i} \ln S_i(t_0))^2}.
\]

Next, we compute the 95% confidence interval for the exponents \( \alpha \) and \( \beta \). For this we follow the book of Montgomery and Peck (15). The 95% confidence interval for \( \beta \) is given by

\[ t_{0.025,N_i - 2} \times se, \]  

where \( t_{0.025,N_i - 2} \) is the \( t \) distribution with parameters \( \alpha/2 \) and \( N_i - 2 \) and \( se \) is the standard error of the exponent defined as

\[
se = \sqrt{SS_{\beta} / (N_i - 2)SS_{xx}},
\]

where \( SS_{\beta} \) is the residual and \( SS_{xx} \) is the variance of \( S_i(t_0) \).

Finally, we express the value of the exponent in terms of the 95% confidence intervals as,

\[ \beta \pm t_{0.025,N_i - 2} \times se. \]

ACKNOWLEDGMENTS. We thank L.H. Dobkins and J. Eckehout for providing data on MSA and C. Briscoe for help with the manuscript. We thank X. Gabaix for useful discussions. This work is supported by National Science Foundation Grant NSF-HSD, and by (J.S.A.). CNpq, CAPES, FUNCAP, and FINEP.

SUPPORTING INFORMATION

Laws of Population Growth

Hernán D. Rozenfeld, Diego Rybski, José S. Andrade Jr.,
Michael Batty, H. Eugene Stanley, and Hernán A. Makse

As supplementary materials we provide the following: In Section I we present tables with
details on our results using the CCA and results presented in previous papers to allow for
comparison between the different approaches. In Section II we study the stability of the
scaling found in the text under a change of scale in the cell size. In Section III we detail the
calculations to relate spatial correlations between the population growth and $\sigma(S_0)$ namely
the relation $\beta = \gamma/4$. In Section IV we describe the random surrogate dataset used to further
test our results. In Section V we further test the robustness of the CCA by proposing a
small variation in the algorithm.

I. CLUSTERS AT DIFFERENT SCALES AND COMPARISON WITH
METROPOLITAN STATISTICAL AREAS

In this section, Tables S1 and S2 allow for a detailed comparison of urban clusters obtained
with the CCA applied to the USA in 1990, and the populations of MSA from US Census
Bureau used in previous studies of population growth [1–3].

We can see that the MSA presented by Eeckhout (2004) typically correspond to our
clusters using cell sizes of 4km and 8km. For example, for the New York City region
Eeckhout’s data are well approximated by a cell size of 4km, but Los Angeles is better
approximated when using a cell size of 8km. On the other hand Dobkins-Ioannides (2000)
data are better described by cell sizes of 2km or 4km. For instance, Chicago is well described
by a cell size of 4km and Los Angeles is better described by a cell size of 2km.

An interesting remark is that the population of Los Angeles when using cell sizes of 2km,
4km and 8km does not vary as much as that for New York. This could be caused by the
fact that major cities in the northeast of USA are closer to each other than large cities in
the southwest, which may be attributed to land or geographical constraints.
It is important relate the results of Table S2 with an ecological fallacy. As the cell size is increased, the population of a cluster also increases, as expected, because the cluster now covers a larger area. This is not a direct manifestation of an ecological fallacy which, would appear if the statistical results (growth rate vs. S or standard deviation vs. S) gave different results as the cell size increases. In Fig. 1 and Fig. 2 in the SI Section II, we observe that the growth rate and standard deviation for the USA and GB follow the same form, except for the case of the growth rate in the USA in which different cell sizes show deviations from each other. The later may be an indicative of an ecological fallacy. In this case, it is not obvious what cell size is the “correct” one. We consider this point (the possibility to choose the cell size) to be a feature of the CCA, since one may appropriately pick the cell size according to the specific problem one is studying.

Table S1: **Top 10 largest MSA of the USA in 1990 from previous analysis of population growth**

<table>
<thead>
<tr>
<th>Dobkins - Ioannides</th>
<th>Eeckhout</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSA</td>
</tr>
<tr>
<td>1</td>
<td>NYC NY206</td>
</tr>
<tr>
<td>2</td>
<td>Los Angeles CA172</td>
</tr>
<tr>
<td>3</td>
<td>Chicago IL59</td>
</tr>
<tr>
<td>4</td>
<td>Philadelphia PA228</td>
</tr>
<tr>
<td>5</td>
<td>Detroit MI80</td>
</tr>
<tr>
<td>6</td>
<td>Washington DC312</td>
</tr>
<tr>
<td>7</td>
<td>San Francisco CA266</td>
</tr>
<tr>
<td>8</td>
<td>Houston TX129</td>
</tr>
<tr>
<td>9</td>
<td>Atlanta GA19</td>
</tr>
<tr>
<td>10</td>
<td>Boston MA39</td>
</tr>
</tbody>
</table>
Table S2: **Top 10 largest clusters of the USA in 1990 from our analysis for different cell sizes.** The city names are the major cities that belong to the clusters and were picked to show the areal extension of the cluster.

<table>
<thead>
<tr>
<th>Cell = 1km</th>
<th>Cell = 2km</th>
<th>Cell = 4km</th>
<th>Cell = 8km</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster</td>
<td>Population</td>
<td>Cluster</td>
<td>Population</td>
</tr>
<tr>
<td>1 NYC</td>
<td>7,012,989</td>
<td>NYC-Long Is.</td>
<td>12,511,237</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Newark</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Jersey City</td>
<td></td>
</tr>
<tr>
<td>2 Chicago</td>
<td>2,312,783</td>
<td>Los Angeles</td>
<td>9,582,507</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Long Beach</td>
<td></td>
</tr>
<tr>
<td>3 Los Angeles</td>
<td>1,411,791</td>
<td>Chicago</td>
<td>4,836,529</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Rockford</td>
<td></td>
</tr>
<tr>
<td>4 Philadelphia</td>
<td>1,282,834</td>
<td>Philadelphia</td>
<td>3,151,704</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Wilmington</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 Boston</td>
<td>759,024</td>
<td>Detroit</td>
<td>2,906,453</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 Newark</td>
<td>581,048</td>
<td>San Francisco</td>
<td>2,601,639</td>
</tr>
<tr>
<td></td>
<td></td>
<td>San Jose</td>
<td></td>
</tr>
<tr>
<td>7 San Francisco</td>
<td>507,300</td>
<td>Washington</td>
<td>2,059,421</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Alexandria</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Bethesda</td>
<td></td>
</tr>
<tr>
<td>8 Washington</td>
<td>504,068</td>
<td>Phoenix</td>
<td>1,556,077</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9 Jersey City</td>
<td>438,591</td>
<td>Boston</td>
<td>1,498,208</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lowell</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Quincy</td>
<td></td>
</tr>
<tr>
<td>10 Baltimore</td>
<td>437,413</td>
<td>Miami</td>
<td>1,465,490</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
FIG. 1: Sensitivity of the results under coarse-graining of the data for GB. (A) Average growth rate and (B) standard deviation for GB using the clustering algorithm for different cell size. The dashed line represents the OLS regression estimate for the exponents (A) $\alpha_{GB} = 0.17$ and (B) $\beta_{GB} = 0.27$ obtained in the main text. For clarity we do not show the confidence bands.

II. SCALING UNDER COARSE-GRAINING

In this section we test the sensitivity of our results to a coarse-graining of the data. We analyze the average growth rate $\langle r(S_0) \rangle$ and the standard deviation $\sigma(S_0)$ for GB and the USA by coarse-graining the data sets at different levels.

In Fig. 1A we observe that although the results are not identical for all coarse-grainings, they are statistically similar, showing a slight decay in the growth rate. Moreover, we see that cities of size $S_0 \approx 10^3$ and $S_0 \approx 10^6$ still exhibit a tendency to have negative growth rates for all levels of coarse-graining, as explained in the main text. In the case of the USA (Fig. 2A) there is a crossover to a flat behavior at a cell size of 8000m, although at this scale all the northeast USA becomes a large cluster of 41 million inhabitants. On the other hand, Figs. 1B, 2B show that the scaling of Eq. (3) in the main text, $\sigma(S_0) \sim S_0^{-\beta}$, still holds when using the coarse-grained datasets on both GB and the USA.

III. CORRELATIONS

In this section we elaborate on the calculations leading to the relation between Gibrat’s law and the spatial correlations in the cell population. We first show that when the pop-
FIG. 2: Study of results under coarse-graining of the data for the USA. (A) Average growth rate and (B) standard deviation for the USA using the clustering algorithm for different cell size. The dashed line represents the OLS regression estimate for the exponents (A) $\alpha_{USA} = 0.28$ and (B) $\beta_{USA} = 0.20$ obtained in the main text. For clarity we do not show the confidence bands.

ulation cells are randomly shuffled (destroying any spatial correlations between the growth rates of the cells), the standard deviation of the growth rate becomes $\sigma(S_0) \sim S_0^{-\beta_{rand}}$, where $\beta_{rand} = 1/2$ [4]. Then, we show that long-range spatial correlations in the population of the cells leads to the relation $\beta = \gamma/4$ as stated at the end of Section II in the main text.

Assuming that the population growth rate is small ($r \ll 1$), we can write $R = e^r \approx 1 + r$. Replacing $R = 1 + r$ in Eq. (1) in the main text we obtain

$$S_1 = S_0 + S_0 r. \tag{1}$$

We define the standard deviation of the populations $S_1$ as $\sigma_1$, which is a function of $S_0$:

$$\sigma_1(S_0) = \sqrt{\langle S_1^2 \rangle - \langle S_1 \rangle^2}. \tag{2}$$

This quantity is easier to relate to the spatial correlations of the cells than the standard deviation $\sigma(S_0)$ of the growth rates $r$. Then, since $\langle S_1 \rangle = S_0 + S_0 \langle r \rangle$ and $\langle S_1^2 \rangle = S_0^2 + 2S_0^2 \langle r \rangle + S_0^2 \langle r^2 \rangle$, we obtain,

$$\sigma_1(S_0) \sim S_0 \sigma(S_0), \tag{3}$$

where $\sigma(S_0) = \sqrt{\langle r^2 \rangle - \langle r \rangle^2}$ as defined in the main text. Therefore, using Eq. (3) in the
As stated in the main text, the total population of a cluster at time $t_0$ is the sum of the populations of each cell, $S_0 = \sum_{j=1}^{N_i} n_j^{(i)}$, where $N_i$ is the number of cells in cluster $i$. The population of a cluster at time $t_1$ can be written as

$$S_1 = S_0 + \sum_{j=1}^{N_i} \delta_j,$$

where $\delta_j$ is the increment in the population of cell $j$ from time $t_0$ to $t_1$ (notice that $\delta_j$ can be negative). Therefore, the standard deviation $\sigma_1(S_0)$ is

$$\left(\sigma_1(S_0)\right)^2 = \sum_{j,k} \langle \delta_j \delta_k \rangle - \langle \sum_j \delta_j \rangle^2 = \sum_{j,k} \langle (\delta_j - \bar{\delta}) (\delta_k - \bar{\delta}) \rangle.$$

(6)

After the process of randomization explained in Section II main text, the correlations between the increment of population in each cell are destroyed. Thus, \newline

$$\langle (\delta_j - \bar{\delta}) (\delta_k - \bar{\delta}) \rangle = \Delta^2 \delta_{jk},$$

(7)

where $\Delta^2 = \bar{\delta}^2 - \bar{\delta}^2$. Replacing in Eq. (6) and since $\langle n \rangle = (1/N_i) \sum_j n_j = S_0/N_i$, we obtain

$$\left(\sigma_1(S_0)\right)^2 = N_i \Delta^2 \sim S_0.$$

(8)

Comparing with Eq. (4) we obtain $\beta_{\text{rand}} = 1/2$ for this uncorrelated case.

Let us assume that the correlation of the population increments $\delta_j$, decays as a power-law of the distance between cells indicating long-range scale-free correlations. Thus, asymptotically

$$\langle (\delta_j - \bar{\delta}) (\delta_k - \bar{\delta}) \rangle \sim \frac{\Delta^2}{|\vec{x}_j - \vec{x}_k|^\gamma},$$

(9)

where $\vec{x}_j$ denotes the position of the cell $j$ and $\gamma$ is the correlation exponent (for $|\vec{x}_j - \vec{x}_k| \to 0$, the correlations $\langle (\delta_j - \bar{\delta}) (\delta_k - \bar{\delta}) \rangle$ tend to a constant). For large clusters, we can approximate the double sum in Eq. (6) by an integral. Then, assuming that the shape of the clusters can be approximated by disks of radius $r_c$, for $\gamma < 2$ we obtain

$$\left(\sigma_1(S_0)\right)^2 = \sum_{j,k} \frac{\Delta^2}{|\vec{x}_j - \vec{x}_k|^\gamma} \approx \frac{\Delta^2}{\alpha^2} \frac{N_i}{r_c} r_c^{-\gamma+2},$$

(10)
where \( a^2 \) is the area of each cell and \( r_c \) the radius of the cluster. Since \( r_c \sim N_i a^2 \), we finally obtain,
\[
\left( \sigma_1(S_0) \right)^2 \sim N_i^{2-\frac{\gamma}{4}}. \tag{11}
\]

Using \( S_0 = N_i \langle n \rangle \) and Eq. (4) we arrive at,
\[
\beta = \frac{\gamma}{4}. \tag{12}
\]

Equation (12) shows that Gibrat’s Law is recovered when the correlation of the population increments is a constant, independent from the positions of the cells; that is when all the populations cells are increased equally. In other words, if \( \gamma = 0 \), the standard deviation of the populations growth rates has no dependence on the population size (\( \beta = 0 \)), as stated by Gibrat’s law. The random case is obtained for \( \gamma = d \), where \( d = 2 \) is the dimensionality of the substrate. In this case \( d = 2 \) and \( \beta_{\text{rand}} = 1/2 \). For \( \gamma > 2 \), the correlations become irrelevant and we still find the uncorrelated case \( \beta_{\text{rand}} = 1/2 \). For intermediate values \( 0 < \gamma < 2 \) we obtain \( 0 < \beta = \gamma/4 < 1/2 \).

**IV. RANDOM SURROGATE DATASET**

In this section we elaborate on the randomization procedure used to understand the role of correlations in population growth.

Figure 4C in the main text shows the standard deviation \( \sigma(S_0) \) when the population of each cluster is randomized, breaking any spatial correlation in population growth. For clusters with a large population, \( \sigma(S_0) \) follows a power-law with exponent \( \beta_{\text{rand}} = 1/2 \), and for small \( S_0 \), \( \sigma(S_0) \) presents deviations from the power-law function as seen in Fig. 4C with smaller standard deviation than the prediction of the random case. This deviation is caused by the fact that the population of a cluster is bound to be positive: a cluster with a small population \( S_0 \) cannot decrease its population by a large number, since it would lead to negative values of \( S_1 \). This produces an upper bound in fluctuations of the growth rate for small \( S_0 \) and results in smaller values of \( \sigma(S_0) \) than expected (below the scaling with exponent \( \beta_{\text{rand}} = 1/2 \)).

To support this argument, we carry out simulations using the clusters of GB, where the population \( n_j(t_0) \) of each cell \( j \) is replaced with random numbers following an exponential distribution with probability \( P(n_j) \sim e^{-n_j/n_0} \). The decay-constant, \( n_0 = 150 \), is extracted
from the data of GB to mimic the original distribution. This is done through a direct measure of \( P(n_j) \) from the GB dataset and fitting the data using OLS regression analysis. We obtain the population \( n_j(t_1) = n_j(t_0) + \delta_j \) of cell \( j \) at time \( t_1 \) by picking random numbers for the population increments \( \delta_j \) following a uniform distribution in the range \(-q \times 150 < \delta_i < q \times 150\). Here \( q \) determines the variance of the increments. Since the population cannot be negative we impose the additional condition \( n_j(t_1) \geq 0 \). Figure 3 shows the results of the standard deviation \( \sigma(S_0) \) for four different \( q \)-values for this uncorrelated model. We find that the tail of \( \sigma(S_0) \) reproduces the uncorrelated exponent \( \beta_{\text{rand}} = 1/2 \). For small \( S_0 \) we find that the standard deviation levels off to an approximately constant value as in the surrogate data of Fig 4C. The crossover from an approximately constant \( \sigma(S_0) \) to a power-law moves to smaller values of the population \( S_0 \) as the standard deviation in the \( \delta_j \) is smaller (smaller value of \( q \)). Such behavior can be understood since the condition \( n_j^{(i)}(t_1) \geq 0 \) imposes a lower “wall” in the random walk specified by \( n_j^{(i)}(t_1) = n_j^{(i)}(t_0) + \delta_j \). As the initial population gets smaller, the walker “feels” the presence of the wall and the fluctuations decrease accordingly, thus explaining the deviations from the power-law with exponent \( \beta_{\text{rand}} = 1/2 \) for small population values. Therefore, as the value of \( q \) decreases, the small population plateau disappears as observed in Fig. 3.

V. A VARIATION OF THE CCA

In this section we study a variation of the CCA. In the main text we stop growing a cluster when the population of all boundary cells have unpopulated, that is, have population exactly 0. In other words, clusters are composed by cell with population strictly greater than 0. It is important to analyze whether this stopping criteria can be relaxed to including cell which have a population larger that a given threshold. In Fig. 4A and Fig. 4B we show the results for the population growth rate and standard deviation, respectively, in GB when the cell size is 2.2km-by-2.2km (as in the main text) but including cells with a population strictly larger than 5 and 20.

Although for small population clusters we observe a slight variation in the growth rate and in the standard deviation, the results show that the thresholds do not influence the global statistics when compared to the plots in the main text.
FIG. 3: Standard deviation $\sigma(S_0)$ for the random data set as explained in the SI Section IV. The results for $\sigma(S_0)$ are rescaled to collapse the power-law tails with exponent $\beta_{\text{rand}} = 1/2$ and to emphasize the deviations from this function for small values of $S_0$. The larger the parameter $q$, the larger the deviations from the power-law at lower $S_0$. In other words, the crossover to power-law tail appears at larger $S_0$ as $q$ increases.

FIG. 4: Sensitivity of the results under a change in the stopping criteria in the CCA (A) Average growth rate for GB with a population threshold of 5 (green line) and 20 (black dashed line) and (B) standard deviation for GB with a population threshold of 5 (green line) and 20 (black dashed line). For clarity we do not show the confidence bands.

