Laws of population growth

Hernán D. Rozenfeld[a], Diego Rybski[a], José S. Andrade, Jr.[b], Michael Batty[c], H. Eugene Stanley[d], and Hernán A. Makse[a,b,1]

[a]Levich Institute and Physics Department, City College of New York, New York, NY 10031; [b]Departamento de Física, Universidade Federal do Ceará, 60451-970 Fortaleza, Ceará, Brazil; [c]Centre for Advanced Spatial Analysis, University College London, 1-19 Torrington Place, London WC1E 6BT, United Kingdom; and [d]Center 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 growth rate 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.

scaling | statistical analysis | urban growth

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 socioeconomical 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, $n_i(t)$, 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 200 m by 200 m (8). At the intermediate scale, we study the USA (continental United States without Alaska), a single country nearly continental in scale, with a population of 303 million in 2007, and an area of 7.44 million km^2. The original USA data consists of 59,456 sites defined by Federal Information Processing Standards (FIPS) accociated 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*A* illustrates operation of the CCA. To identify urban clusters, we require connected cells to have nonzero population. We

Author contributions: M.B., H.E.S., and H.A.M. designed research; H.D.R., D.R., J.S.A., and H.A.M. performed research; H.D.R., D.R., J.S.A., and H.A.M. analyzed data; and H.D.R., D.R., J.S.A., M.B., and H.A.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence should be addressed. E-mail: hmakse@lev.ccny.cuny.edu.

This article contains supporting information online at www.pnas.org/cgi/content/full/0807435105/DCSupplemental.

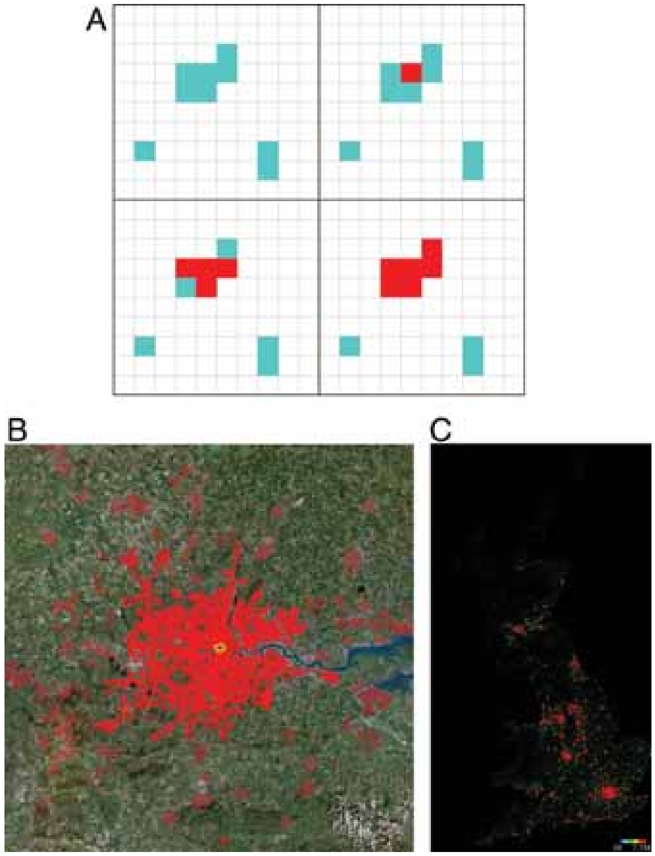

Fig. 1. (*A*) Illustration of the CCA applied to a sample of gridded population data. In *Upper Left*, cells are blue if they are populated ($n_j^{(i)}(t) > 0$), otherwise, if $n_j^{(i)}(t) = 0$, they are white. In *Upper Right*, we initialize the CCA by selecting a populated cell and burning it (red cell). Then, we burn the populated neighbors of the red cell as shown in *Lower Left*. We keep growing the cluster by iteratively burning neighbors of the red cells until all neighboring cells are unpopulated, as shown in *Lower Right*. Next, we pick another unburned populated cell and repeat the algorithm until all populated cells are assigned to a cluster. The population $S_i(t)$ of cluster i at time t is then $S_i(t) = \sum_{j=1}^{N_i} n_j^{(i)}(t)$. (*B*) Cluster identified with the CCA in the London area (red) overlaying a corresponding satellite image (extracted from maps.google.com). The greenery corresponds to vegetation, and thus approximately indicates unoccupied areas. For example, Richmond Park can be found as a vegetation area in the southwest. The areas in the east along the Thames River correspond mainly to industrial districts, and in the west to the London Heathrow Airport, also not populated. The yellow line in the center represents the administrative boundary of the City of London, demonstrating the difference with the urban cluster found with the CCA. The pink clusters surrounding the major red cluster are smaller conglomerates not connected to London. The figure shows that an analysis based on the City of London captures only a partial area of the real urban agglomeration. (*C*) Result of the CCA applied to all of Great Britain showing the large variability in the population distribution. The color bar (in logarithmic scale) indicates the population of each urban cluster.

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 $n_j^{(i)}(t)$ of each cell j within it, $S_i(t) = \sum_{j=1}^{N_i} n_j^{(i)}(t)$, where N_i is the number of cells in the cluster. Results of the CCA are shown in Fig. 1*B*, 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. 1*B*). Fig. 1*C* 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

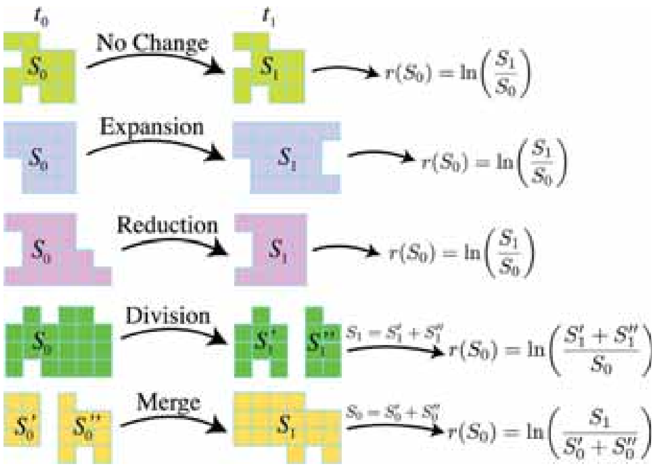
t_0
t_1
No Change
S_0
S_1
$r(S_0) = \ln\left(\frac{S_1}{S_0}\right)$
Expansion
S_0
S_1
$r(S_0) = \ln\left(\frac{S_1}{S_0}\right)$
Reduction
S_0
S_1
$r(S_0) = \ln\left(\frac{S_1}{S_0}\right)$
Division
S_0
S_1'
S_1''
$S_1 = S_1' + S_1''$
$r(S_0) = \ln\left(\frac{S_1' + S_1''}{S_0}\right)$
Merge
S_0'
S_0''
S_1
$S_0 = S_0' + S_0''$
$r(S_0) = \ln\left(\frac{S_1}{S_0' + S_0''}\right)$

Fig. 2. Illustration of possible changes in cluster shapes. In each case we show how the growth rate is computed. In the first case, there is no areal modification in the cluster between t_0 and t_1. In the second, the cluster expands. In the third, the cluster reduces its area. In the fourth, one cluster divides into two and therefore we consider the population at t_1 to be $S_1 = S_1' + S_1''$. In the fifth case, two clusters merge to form one at t_1. In this case, we consider the population at t_0 to be $S_0 = S_0' + S_0''$.

the populations of a given city or cluster i, at times t_0 and $t_1 > t_0$, are related by

$$S_1 = R(S_0)S_0, \qquad [1]$$

where $S_0 \equiv S_i(t_0) = \sum_j^{N_i} n_j^{(i)}(t_0)$ and $S_1 \equiv S_i(t_1) = \sum_j^{N_i} n_j^{(i)}(t_1)$ 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 $r(S_0) \equiv \ln R(S_0) = \ln(S_1/S_0)$, and study the dependence of the mean value of the growth rate, $\langle r(S_0)\rangle$, and the standard deviation, $\sigma(S_0) = \sqrt{\langle r(S_0)^2\rangle - \langle r(S_0)\rangle^2}$, on the initial population, S_0. The averages $\langle r(S_0)\rangle$ and $\sigma(S_0)$ are calculated by applying nonparametric techniques (13, 14) (see *Calculation of* $\langle r(S_0)\rangle$ *and* $\sigma(S_0)$ *and Methodology* for details). To obtain the population growth rate of clusters we take into account that not all clusters occupy the same area between t_0 and t_1 according to the cases discussed in Fig. 2. The figure shows how to calculate the growth rate $r(S_0)$ in each case.

We analyze the population growth in the United States from $t_0 = 1990$ to $t_1 = 2000$ (9). We apply the CCA to identify the clusters in the data of 1990 and calculate their growth rates by comparing them with the population of the same clusters in 2000 when the data are gridded with a cell size of 2 km by 2 km. We calculate the annual growth rates by dividing r by the time interval $t_1 - t_0$.

Fig. 3*A* shows a nonparametric regression with bootstrapped 95% confidence bands (13, 14) of the growth rate of the USA, $\langle r(S_0)\rangle$ (see *Calculation of* $\langle r(S_0)\rangle$ *and* $\sigma(S_0)$ *and Methodology* for details). We find that the growth rate diminishes from $\langle r(S_0)\rangle \approx 0.012 \pm 0.004$ (error includes the confidence bands) for populations $<10^4$ inhabitants to $\langle r(S_0)\rangle \approx 0.002 \pm 0.002$ for the largest populations at approximately $S_0 \approx 10^7$. We may argue that the mean growth rate deviates from Gibrat's law beyond the confidence bands. Although it is difficult to fit the data to a single function for the entire range, the data show a decrease with S_0 approximately after a power law in the tail for populations $>10^4$. An attempt to fit the data with a power law yields the following scaling in the tail:

$$\langle r(S_0)\rangle \sim S_0^{-\alpha}, \qquad [2]$$

where α is the mean growth exponent, which takes a value $\alpha_{\mathrm{USA}} = 0.28 \pm 0.08$ from Ordinary Least Squares (OLS) analysis (15) (see *Calculation of* $\langle r(S_0)\rangle$ *and* $\sigma(S_0)$ *and Methodology* for details on OLS and on the estimation of the exponent error).

Fig. 3*B* 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,

$$\sigma(S_0) \sim S_0^{-\beta}, \qquad [3]$$

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

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 2.2 km by 2.2 km at $t_0 = 1981$ and $t_1 = 1991$. The average growth rate depicted in Fig. 4*A* 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 $S_0 \approx 10^4$ 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 $S_0 = 10^3$ exhibit negative growth rates as well. Thus, the mean rates are plotted on a semilogarithmic scale in Fig. 4*A*. 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 $\alpha_{\mathrm{GB}} = 0.17 \pm 0.05$ from OLS regression analysis, as shown in Fig. 4*A Inset*. Fig. 4*B* 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, $\beta_{\mathrm{GB}} = 0.27 \pm 0.04$, obtained with OLS technique.

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.

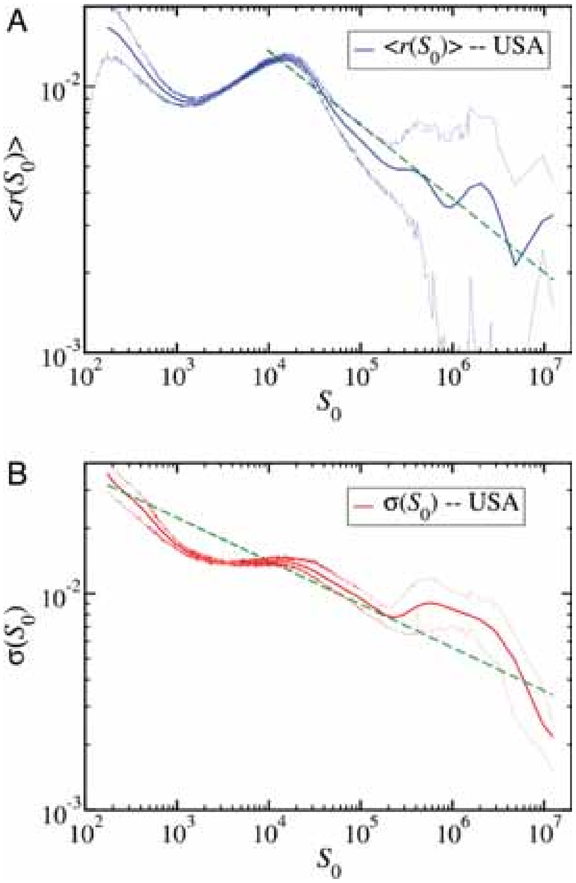

Fig. 3. Results for the United States by using a cell size of 2 km by 2 km. (*A*) Mean annual growth rate for population clusters in the Unites States versus the initial population of the clusters. The straight dashed line shows a power-law fit with $\alpha_{\mathrm{USA}} = 0.28 \pm 0.08$ as determined by using OLS regression. (*B*) Standard deviation of the growth rate for the United States. The straight dashed line corresponds to a power-law fit using OLS regression with $\beta_{\mathrm{USA}} = 0.20 \pm 0.06$.

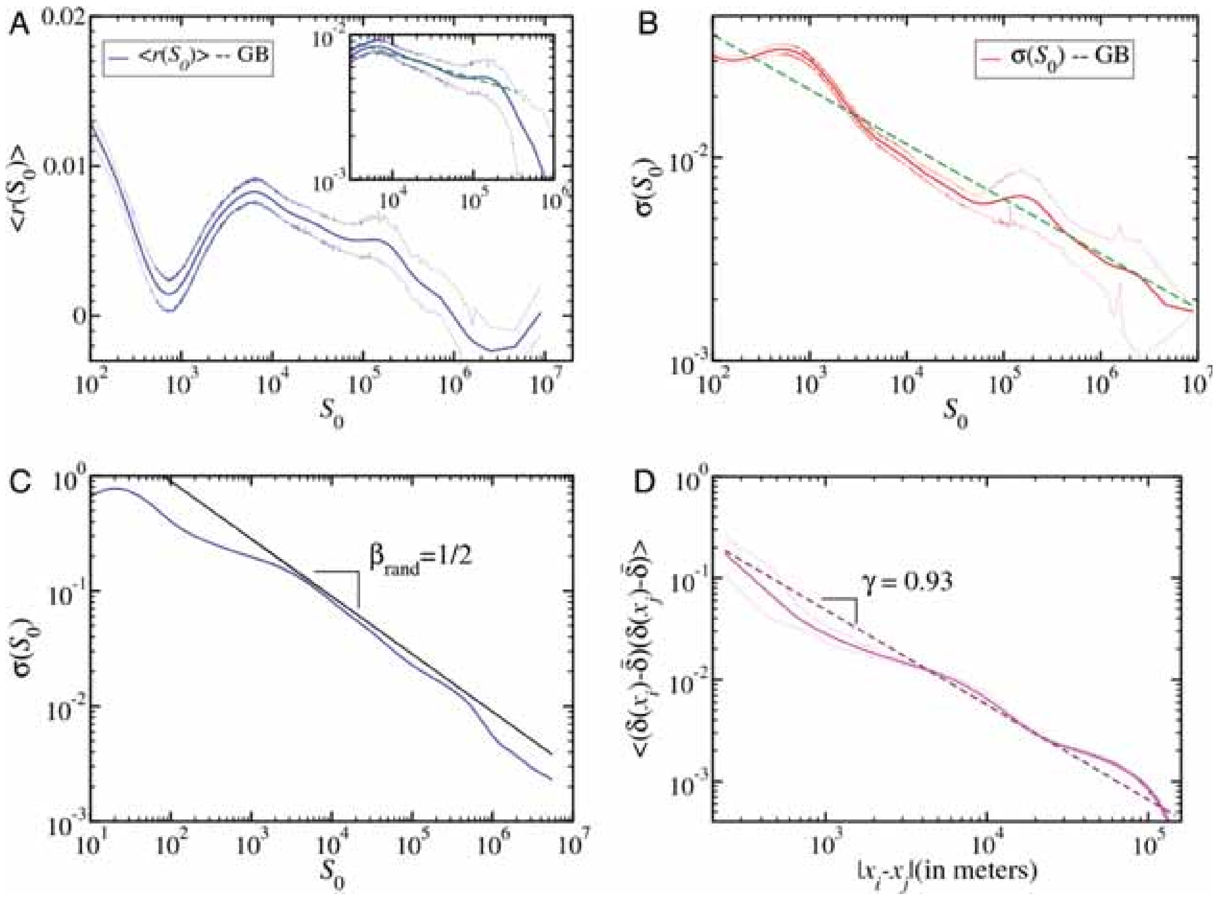

Fig. 4. Results for Great Britain by using a cell size of 2.2 km by 2.2 km. (*A*) Mean annual growth rate of population clusters in Great Britain versus the initial cluster population. *Inset* Double-logarithmic plot of the growth rate in the intermediate range of populations, $3,000 < S_0 < 3 \times 10^5$. A power-law fit using OLS leads to an exponent $\alpha_{\mathrm{GB}} = 0.17 \pm 0.05$ for this range. (*B*) Double-logarithmic plot of the standard deviation of the annual growth rates of population clusters in Great Britain versus the initial cluster population. The straight line corresponds to a power-law fit using OLS with an exponent $\beta_{\mathrm{GB}} = 0.27 \pm 0.04$, according to Eq. **3**. (*C*) Scaling of the standard deviation in cluster population obtained from the randomized surrogate dataset of Great Britain by randomly swapping the cells. The data show an exponent $\beta_{\mathrm{rand}} = 1/2$ in the tail. The deviations for small S_0 are discussed in the *SI Appendix*, Section IV, where we test these results by generating random populations. (*D*) Long-range spatial correlations in the population growth of cells for Great Britain according to Eq. **6**. The straight line corresponds to an exponent $\gamma = 0.93 \pm 0.08$.

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. 5*A* 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. 5*A* reveals a power-law scaling $\langle r(S_0)\rangle \sim S_0^{-\alpha_{\mathrm{Af}}}$, with $\alpha_{\mathrm{Af}} = 0.21 \pm 0.05$ from OLS regression analysis. The standard deviation (Fig. 5*B*) satisfies Eq. **3** with a standard deviation exponent $\beta_{\mathrm{Af}} = 0.19 \pm 0.04$.

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, $n_j^{(i)}(t_0)$ and $n_k^{(i)}(t_0)$, at time t_0. This swapping process preserves the probability distribution of $n_j^{(i)}$, but destroys any spatial correlations among the population cells. Fig. 4*C* 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 $\beta_{\mathrm{rand}} = 1/2$. 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 δ_j with mean value $\bar{\delta}$ and variance $\langle(\delta - \bar{\delta})^2\rangle = \Delta^2$, and that $r \ll 1$, then $n_j^{(i)}(t_1) = n_j^{(i)}(t_0) + \delta_j$. Therefore, the population of a cluster at time t_1 can be written as

$$S_1 = S_0 + \sum_{j=1}^{N_i} \delta_j. \qquad [4]$$

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

$$\langle S_1^2\rangle = \langle S_0^2\rangle + \sum_{j}^{N_i} \sum_{k}^{N_i} \langle(\delta_j - \bar{\delta})(\delta_k - \bar{\delta})\rangle. \qquad [5]$$

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

The fact that β lies below the random exponent ($\beta_{\mathrm{rand}} = 1/2$) for all the analyzed data suggests that the dynamics of the population cells display spatial correlations, which are eliminated in

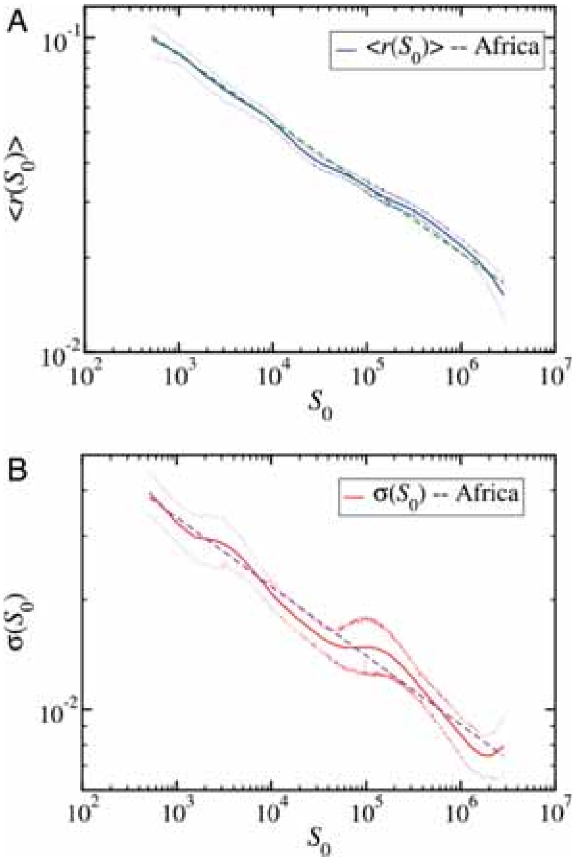

Fig. 5. Results for Africa by using a cell size of 77 km by 77 km. (*A*) Mean growth rate of clusters in Africa versus the initial size of population S_0. The straight dashed line shows a power-law fit with exponent $\alpha_{\mathrm{Af}} = 0.21 \pm 0.05$, obtained by using OLS regression. (*B*) Standard deviation of the growth rate in Africa. The straight line corresponds to power-law fit by using OLS providing the exponent $\beta_{\mathrm{Af}} = 0.19 \pm 0.04$.

the random surrogate data. 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, δ_j, are asymptotically of a scale-invariant form characterized by a correlation exponent γ,

$$\langle(\delta_j - \bar{\delta})(\delta_k - \bar{\delta})\rangle \sim \frac{\delta^2}{|\vec{x}_j - \vec{x}_k|^{\gamma}}, \qquad [6]$$

where $\vec{x}_j$ is the location of cell j. For Great Britain we find $\gamma = 0.93 \pm 0.08$ (see Fig. 4*D*). 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 $\beta_{\mathrm{rand}} = 1/2$ (larger values of γ result in the same β 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 $\beta_{\mathrm{GB}} = 0.27 \pm 0.04$, 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/~hmakse/cities/city_data.zip.

The datasets consist of a list of populations at specific coordinates at 2 time steps, t_0 and t_1. A graphical representation of the data can be seen in Fig. 1*C* 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 $\langle r(S_0)\rangle$ 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 $P(r|S_0)$ 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,

$$\langle r(S_0)\rangle = \int rP(r|S_0)\mathrm{d}r, \qquad [7]$$

Table 1. Characteristics of datasets and summary of results

Data	No. of cells	t_0	t_1	Average growth rate, %	Cell size	No. of clusters	α	β
USA	1.86 mill	1990	2000	0.9	2 km by 2 km	30,210	0.28 ± 0.08	0.20 ± 0.06
GB	0.10 mill	1981	1991	0.3	2.2 km by 2.2 km	10,178	0.17 ± 0.05	0.27 ± 0.04
Africa	2,216	1960	1990	4	77 km by 77 km	3,988	0.21 ± 0.05	0.19 ± 0.04

and

$$\langle r(S_0)^2 \rangle = \int r^2 P(r|S_0) \mathrm{d}r. \qquad [8]$$

Once $r(S_0)$ and $\sigma(S_0)$ 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_0 and between the standard deviation and S_0. Following the methods explained in ref. 14, we apply the Nadaraya–Watson method to calculate an estimate for the growth rate, $\hat{r}(S_0)$, with,

$$\langle \hat{r}(S_0) \rangle = \frac{\sum_{i=0}^{\text{allclusters}} K_h(S_0 - S_i(t_0)) r_i(S_0)}{\sum_{i=0}^{\text{allclusters}} K_h(S_0 - S_i(t_0))}, \qquad [9]$$

and an estimate for the standard deviation $\hat{\sigma}(S_0)$ with,

$$\hat{\sigma}(S_0) = \sqrt{\frac{\sum_{i=0}^{\text{allclusters}} K_h(S_0 - S_i(t_0))(r_i(S_0) - \langle \hat{r}(S_0) \rangle)^2}{\sum_{i=0}^{\text{allclusters}} K_h(S_0 - S_i(t_0))}}, \qquad [10]$$

where $S_i(t_0)$ is the population of cluster i at time t_0 (as defined in the main text), $r_i(S_0)$ is the growth rate of cluster i, and $K_h(S_0 - S_i(t_0))$ is a Gaussian kernel of the form,

$$K_h(S_0 - S_i(t_0)) = \mathrm{e}^{\frac{(\ln S_0 - \ln S_i(t_0))^2}{2h^2}}, \quad h = 0.5 \qquad [11]$$

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 $\alpha/2$ (α is not related to the growth rate exponent) quantile we obtain the 95% confidence bands.

To obtain the exponents α and β of the power-law scalings for $\langle r(S_0) \rangle$ and $\sigma(S_0)$, respectively, we perform an OLS regression analysis (15). More specifically, to obtain the exponent β from Eq. **3**, we first linearize the data by considering the logarithm of the independent and dependent variables so that Eq. **3** becomes $\ln\sigma(S_0) \sim \beta \ln S_0$. Then, we apply a linear OLS regression that leads to the exponent

$$\beta = \frac{N_c \sum_{i=1}^{N_c} [\ln S_i(t_0) \ln \sigma(S_i(t_0))] - \sum_{i=1}^{N_c} \ln S_i(t_0) \sum_{i=1}^{N_c} \ln \sigma(S_i(t_0))}{N_c \sum_{i=1}^{N_c} (\ln S_i(t_0))^2 - \left(\sum_{i=1}^{N_c} \ln S_i(t_0) \right)^2}, \qquad [12]$$

where N_c is the number of clusters found by using the CCA. Analogously, we obtain the exponent α by linearizing $\langle |r(S_0)| \rangle$ and calculating

$$\alpha = \frac{N_c \sum_{i=1}^{N_c} (\ln S_i(t_0) \ln \langle |r(S_i(t_0))| \rangle - \sum_{i=1}^{N_c} \ln S_i(t_0) \sum_{i=1}^{N_c} \ln \langle |r(S_i(t_0))| \rangle}{N_c \sum_{i=1}^{N_c} (\ln S_i(t_0))^2 - (\sum_{i=1}^{N_c} \ln S_i(t_0))^2}. \qquad [13]$$

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

$$t_{0.025, N_c - 2} * se, \qquad [14]$$

where $t_{\alpha'/2, N_c - 2}$ is the t distribution with parameters $\alpha'/2$ and $N_c - 2$ and se is the standard error of the exponent defined as

$$se = \sqrt{\frac{SS_E}{(N_c - 2) S_{xx}}}, \qquad [15]$$

where SS_E is the residual and S_{xx} is the variance of S_0.

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

$$\beta \pm t_{0.025, N_c - 2} * se. \qquad [16]$$

ACKNOWLEDGMENTS. We thank L.H. Dobkins and J. Eeckhout 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.

1. Gabaix X (1999) Zipf's law for cities: an explanation. *Q J Econ* 114:739–767.
2. Gabaix X, Ioannides YM (2003) The evolution of city size distributions. *Handbook of Urban and Regional Economics*, eds Henderson JV, Thisse JF (Elsevier Science, Amsterdam) Vol 4, pp 2341–2378.
3. Unwin DJ (1996) GIS, spatial analysis and spatial statistics. *Progr Hum Geogr* 20:540–551.
4. King G, Rosen O, Tanner MA, eds (2004) *Ecological Inference: New Methodological Strategies* (Cambridge Univ Press, New York).
5. Eeckhout J (2004) Gibrat's law for (all) cities. *Am Econ Rev* 94:1429–1451.
6. Dobkins LH, Ioannides YM (2000) Spatial interactions among U.S. cities: 1900–1990. *Reg Sci Urban Econ* 31:701–731.
7. Ioannides YM, Overman HG (2003) Zipf's law for cities: An empirical examination. *Reg Sci Urban Econ* 33:127–137.
8. The 1981 and 1991 population census, Crown Copyright, ESRC purchase. Available at: http://census.ac.uk/. Last accessed: July 30, 2008.
9. Environmental Systems Research Institute (ESRI) (2000) ArcView 3.2 data sets: North America (ESRI, Redlands, CA).
10. United Nations Environment Programme/Global Resource Information Database (UNEP/GRID) (1987) through GRID. Available at: http://na.unep.net/datasets/datalist.php. Last accessed: July 30, 2008.
11. Stauffer D (1984) *Introduction to Percolation Theory* (Taylor & Francis, London).
12. Eaton J, Eckstein Z (1997) Cities and growth: Theory and evidence from France and Japan. *Reg Sci Urban Econ* 27:443–474.
13. Härdle W (1990) *Applied Nonparametric Regression* (Cambridge Univ Press, Cambridge).
14. Silverman BW (1986) *Density Estimation for Statistics and Data Analysis* (Chapman and Hall, New York).
15. Montogomery DC, Peck EA (1992) *Introduction to Linear Regression Analysis* (Wiley, New York).
16. Stanley MHR, *et al.* (1996) Scaling behavior in the growth of companies. *Nature* 379:804–806.
17. Makse HA, Havlin S, Stanley HE (1995) Modelling urban growth patterns. *Nature* 377:608–612.
18. Barabási A-L, Albert R (1999) Emergence of scaling in random networks. *Science* 286:509–512.
19. Carlson JM, Doyle J (2002) Complexity and robustness. *Proc Natl Acad Sci USA* 99:2538–2545.
20. Rossi-Hansberg E, Wright MLJ (2007) Establishment size dynamics in the aggregate economy. *Am Econ Rev* 97:1639–1666.