## 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**

| | Dobkins - Ioannides | | Eeckhout | |
|---|---|---|---|---|
| | MSA | Population | MSA | Population |
| 1 | NYC NY206 | 9,372,000 | NYC-North NJ-Long Is., NY-NJ-CT-PA | 19,549,649 |
| 2 | Los Angeles CA172 | 8,863,000 | Los Angeles-Riverside-Orange County, CA | 14,531,529 |
| 3 | Chicago IL59 | 7,333,000 | Chicago-Gary-Kenosha, IL-IN-WI | 8,239,820 |
| 4 | Philadelphia PA228 | 4,857,000 | Washington-Baltimore, DC-MD-VA-WV | 6,727,050 |
| 5 | Detroit MI80 | 4,382,000 | San Francisco-Oakland-San Jose, CA | 6,253,311 |
| 6 | Washington DC312 | 3,924,000 | Philadelphia-Wilmington-Atlantic City PA-NJ-DE-MD | 5,892,937 |
| 7 | San Francisco CA266 | 3,687,000 | Boston-Worcester-Lawrence, MA-NH-ME-CT | 5,455,403 |
| 8 | Houston TX129 | 3,494,000 | Detroit-Ann Arbor-Flint, MI | 5,187,171 |
| 9 | Atlanta GA19 | 2,834,000 | Dallas-Fort Worth, TX | 4,037,282 |
| 10 | Boston MA39 | 2,800,000 | Houston-Galveston-Brazoria, TX | 3,731,131 |

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.

| | Cell = 1km | | Cell = 2km | | Cell = 4km | | Cell = 8km | |
|---|---|---|---|---|---|---|---|---|
| | Cluster | Population | Cluster | Population | Cluster | Population | Cluster | Population |
| 1 | NYC | 7,012,989 | NYC-Long Is.<br>Newark<br>Jersey City | 12,511,237 | NYC-Long Is.<br>N. NJ-Newark<br>Jersey City | 17,064,816 | NYC-Long Is.<br>North NJ<br>Philadelphia<br>D.C.-Boston | 41,817,858 |
| 2 | Chicago | 2,312,783 | Los Angeles<br>Long Beach | 9,582,507 | Los Angeles<br>Long Beach<br>Pomona | 10,878,034 | Los Angeles<br>San Clemente<br>Riverside | 13,304,233 |
| 3 | Los Angeles | 1,411,791 | Chicago<br>Rockford | 4,836,529 | Chicago<br>Gary<br>Rockford | 7,230,404 | Chicago<br>Gary<br>Rockford<br>Milwaukee | 9,288,345 |
| 4 | Philadelphia | 1,282,834 | Philadelphia<br>Wilmington | 3,151,704 | Washington<br>Baltimore<br>Springfield | 5,316,890 | San Francisco<br>Santa Cruz<br>Brentwood | 5,736,479 |
| 5 | Boston | 759,024 | Detroit | 2,906,453 | Philadelphia<br>Trenton<br>Wilmington | 4,935,734 | Detroit<br>Ann Arbor<br>Monroe<br>Sarnia | 4,442,723 |
| 6 | Newark | 581,048 | San Francisco<br>San Jose | 2,601,639 | San Francisco<br>San Jose<br>Concord | 4,766,960 | Miami<br>Port St. Lucie | 4,000,432 |
| 7 | San Francisco | 507,300 | Washington<br>Alexandria<br>Bethesda | 2,059,421 | Detroit<br>Waterford<br>Canton | 3,722,778 | Dallas<br>Fort Worth | 3,536,186 |
| 8 | Washington | 504,068 | Phoenix | 1,556,077 | Miami<br>W. Palm Beach | 3,719,773 | Houston | 3,425,647 |
| 9 | Jersey City | 438,591 | Boston<br>Lowell<br>Quincy | 1,498,208 | Dallas<br>Fort Worth | 3,134,233 | Cleveland<br>Canton | 3,233,341 |
| 10 | Baltimore | 437,413 | Miami | 1,465,490 | Boston<br>Brockton<br>Nashua | 3,064,925 | Pittsburgh<br>Youngstown<br>Morgantown | 3,214,661 |

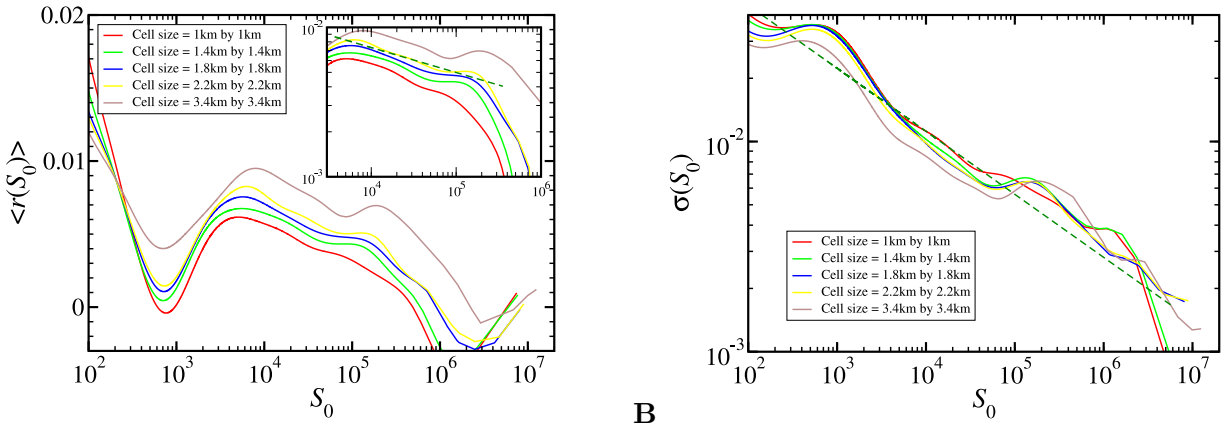


FIG. 1: Sensitivity of the results under coarse-graining of the data for GB. **(A)** Average growth rate and **(B)** standard deviation for GB using the clustering algorithm for different cell size. The dashed line represents the OLS regression estimate for the exponents **(A)** $\alpha_{\rm GB} = 0.17$ and **(B)** $\beta_{\rm GB} = 0.27$ 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-

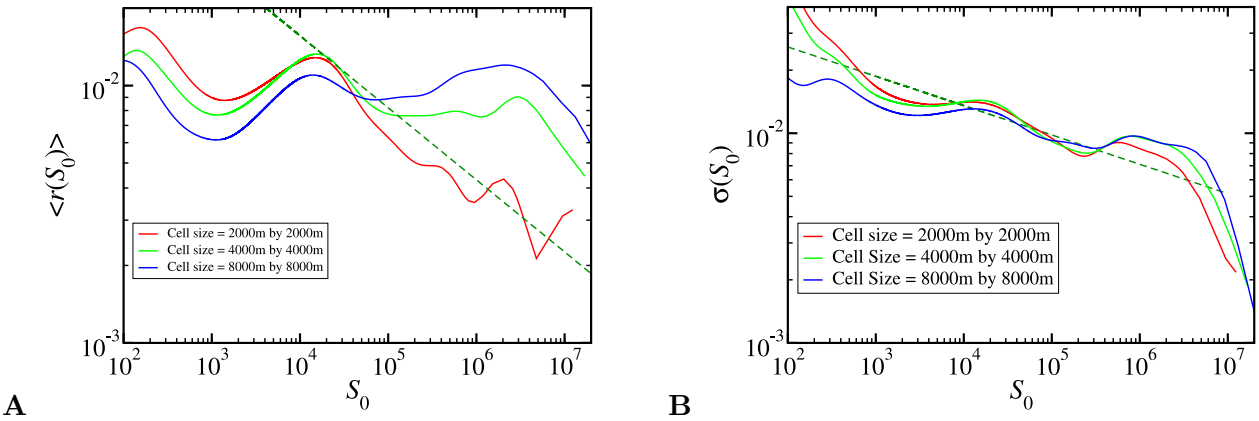


FIG. 2: Study of results under coarse-graining of the data for the USA. **(A)** Average growth rate and **(B)** standard deviation for the USA using the clustering algorithm for different cell size. The dashed line represents the OLS regression estimate for the exponents **(A)** $\alpha_{\rm USA} = 0.28$ and **(B)** $\beta_{\rm USA} = 0.20$ 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 $\sigma(S_0) \sim S_0^{-\beta_{\rm rand}}$, where $\beta_{\rm rand} = 1/2$ [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

main text,

$$\sigma_1(S_0) \sim S_0^{1-\beta}. \tag{4}$$

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

$$S_1 = S_0 + \sum_{j=1}^{N_i} \delta_j, \tag{5}$$

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

$$\Big(\sigma_1(S_0)\Big)^2 = \sum_{j,k}^{N_i} \langle \delta_j \delta_k \rangle - \langle \sum_{j}^{N_i} \delta_j \rangle^2 = \sum_{j,k}^{N_i} \langle (\delta_j - \bar{\delta})(\delta_k - \bar{\delta}) \rangle. \tag{6}$$

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

$$\langle (\delta_j - \bar{\delta})(\delta_k - \bar{\delta}) \rangle = \Delta^2 \delta_{jk}, \tag{7}$$

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

$$\Big(\sigma_1(S_0)\Big)^2 = N_i \Delta^2 \sim S_0. \tag{8}$$

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

$$\langle (\delta_j - \bar{\delta})(\delta_k - \bar{\delta}) \rangle \sim \frac{\Delta^2}{|\vec{x}_j - \vec{x}_k|^\gamma}, \tag{9}$$

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

$$(\sigma_1(S_0))^2 = \sum_{j,k}^{N_i} \frac{\Delta^2}{|\vec{x}_j - \vec{x}_k|^\gamma} \to \Delta^2 \frac{N_i}{a^2} \int^{r_c} \frac{r \mathrm{d}r \mathrm{d}\theta}{r^\gamma} \approx \frac{\Delta^2}{(2-\gamma)} \frac{N_i}{a^2} r_c^{-\gamma+2}, \tag{10}$$

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,

$$\Big(\sigma_1(S_0)\Big)^2 \sim N_i^{2-\frac{\gamma}{2}}. \tag{11}$$

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 $P(n_j) \sim \mathrm{e}^{-n_j/n_0}$. 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 $-q*150 < \delta_i < q*150$. 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 $\beta_{\rm rand} = 1/2$. 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 $\beta_{\rm rand} = 1/2$ 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.

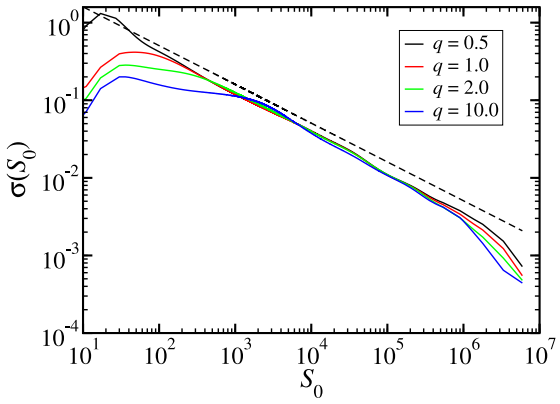


FIG. 3: Standard deviation $\sigma(S_0)$ for the random data set as explained in the SI Section IV. The results for $\sigma(S_0)$ are rescaled to collapse the power-law tails with exponent $\beta_{\rm rand} = 1/2$ and to emphasize the deviations from this function for small values of $S_0$. The larger 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.

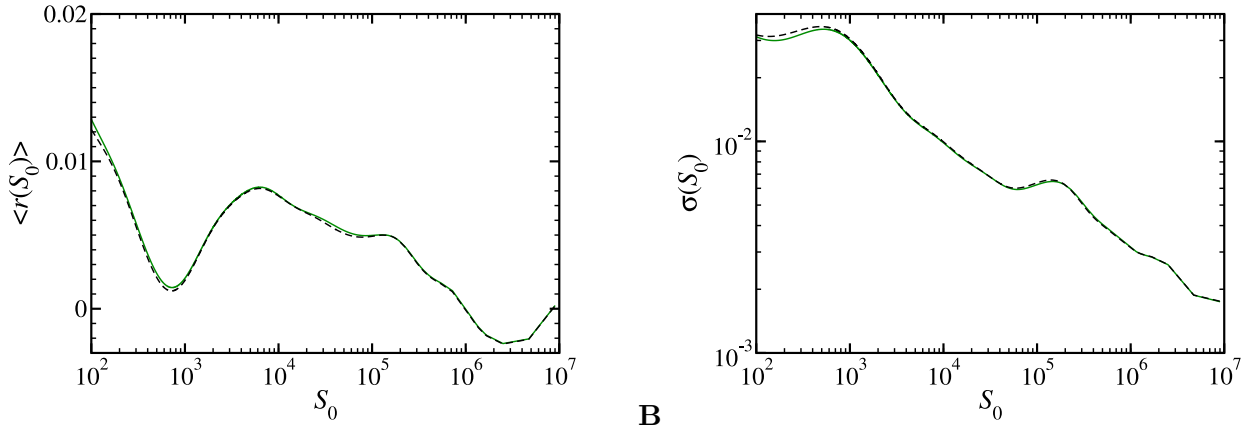


FIG. 4: Sensitivity of the results under a change in the stopping criteria in the CCA **(A)** Average growth rate for GB with a population threshold of 5 (green line) and 20 (black dashed line) and **(B)** standard deviation for GB with a population threshold of 5 (green line) and 20 (black dashed line).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

[1] Eeckhout J (2004) Gibrat's law for (all) cities. *Amer. Econ. Rev.* 94: 1429-1451.

[2] Dobkins L H, Ioannides Y M (2000) Spatial interactions among U.S. cities: 1900-1990. *Reg. Sci. Urban Econ.* 31: 701-731.

[3] Ioannides Y M, Overman H G (2003) Zipf's law for cities: an empirical examination *Reg. Sci. Urban Econ.* 33: 127-137.

[4] Stanley M H R *et al.* (1996) Scaling behavior in the growth of companies. *Nature* 379: 804-806.