# Science Advances

MAAS

advances.sciencemag.org/cgi/content/full/6/34/eaat8812/DC1

## Supplementary Materials for

## Urban growth and the emergent statistics of cities

Luis M. A. Bettencourt*

*Corresponding author. Email: bettencourt@uchicago.edu

**This PDF file includes:**

# 1 Scaling relations and tests of ordinary least squares fitting

Urban indicators, $Y$, such as wages, are typically non-linear stochastic functions of city population size, $N$. Scaling analysis posits that we can identify relationships that are linear in specific parameters, $\beta$, known as exponents, starting from a random variable model for $Y$ of the form

$$\ln Y_i(N_i, t) = \ln Y_0(t) + \beta \ln N_i(t) + \xi_i(t).$$

This model is the logarithmically-transformed scaling relation, Eq. (1). Because this relationship is linear in the parameter $\beta$, its value can now be estimated via *ordinary least squares* (OLS) minimization of the model against data, as shown in Figure 2. For this simple procedure to be justified, standard conditions for the applicability of OLS estimation must apply. Whether such conditions are satisfied is an empirical question, which may depend on the quantity $Y$ itself, on the city definition adopted and the delineation of the urban system (e.g. a nation, or a region).

OLS estimation of the $\beta$ with a simple one-variable model like Eq. (1-2), requires that the residuals (or errors) $\xi_i(t)$ obey a number of well-known conditions:

1. The regression model is linear in the parameters to be estimated and the error term;

2. The residuals' ensemble mean is zero : $\langle \xi(t) \rangle = 0$;

3. Independent scaling variables – here just $\ln N_i(t)$– are uncorrelated with the errors $\xi_i$;

4. Error observations are mutually uncorrelated in the dataset;

5. Errors have a constant variance (homoskedasticity);

6. No perfect correlations between independent variables;

7. The error term, $\xi_i$ is normally distributed (optional).

We now show explicitly that these conditions are satisfied for the scaling analysis of Figure 2. Condition 1 is clearly satisfied by visual inspection of Figures 2A-B, but can also be measured as a very high linear goodness of fit in terms of its $R^2 = 0.935$ (similar across years). Condition 2 is clearly apparent in Figure 2C and 5D; for each and every year, but can be quantified by computing the explicit average $|\langle \xi(t) \rangle| < 10^{-12}$ for all years. Condition 3 is easily assessed by inspection of Figure S1; which as a statistical test shows $R^2 < 10^{-3}$ for the correlation between residuals and $\ln N_i$ in any year. Condition 4 is moot, given that the data are not a sample but the collection of all metropolitan areas in a nation; in this sense there can be no sampling bias. Condition 5 of homoscedasticity is tested in Figure S1 and further discussed below. Condition 6 is also moot given that scaling analysis has a single independent variable, $\ln N_i$. Finally, Condition 7 is indeed the major subject of the paper. We find that the statistics of flows, such as wages, are not strictly normally distributed, but asymptotically well fit by two Gaussians as in Figure 5D. The normality condition is optional for the validity of OLS estimation so long as the distribution of the errors is sufficiently narrow, symmetric and with mean zero, which is indeed the case. The consequence of the (slight) non-normality of flow variables observed means that the 95% confidence intervals quoted in the captions may actually be associated with a slightly larger confidence of including the data than expected from a normal model for the $\xi_i(t)$.

Let us now provide some addition evidence for homoscedasticity (condition 5), see Figure S1. For each year, we divided the ensemble of US Metropolitan Statistical Areas into quartiles (smallest cities, medium small cities, medium large cities, largest cities) and measured the corresponding subset standard deviation of residuals. Figure S1 immediately shows that there is very little to no correlation between the variation in the $\xi_i(t)$ (blue dots) and city size. Further, quantitative estimation of the standard deviation in each quartile (population size quartiles are indicated by dashed vertical red lines) is shown as red dots. There are small variations in the quartile standard deviations, but all these are essentially the same with no consistent variation
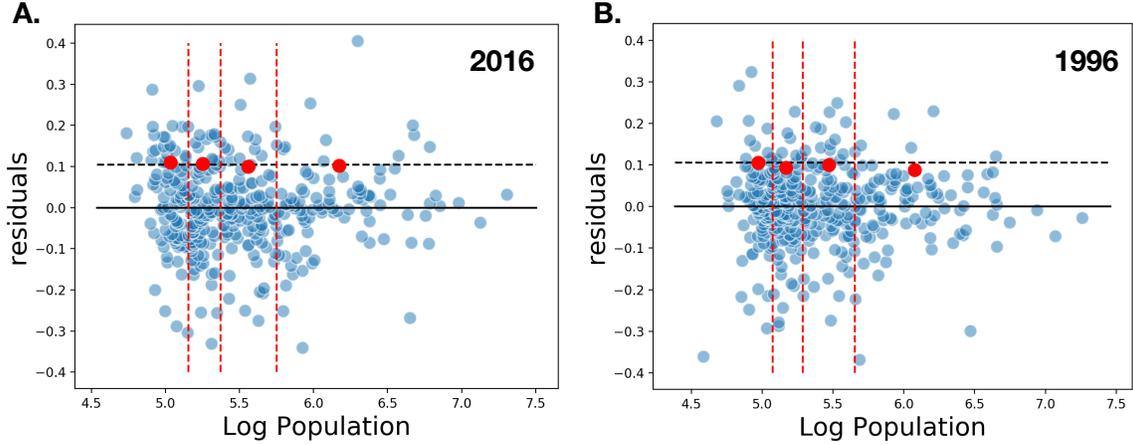
Figure S1: **The residuals of scaling analysis of annual wages in US Metropolitan Statistical Areas, Figures 2A-B, show a city-size independent variance (homoscedasticity)**. Residuals $\xi_i(t)$ for each city (blue dots) are shown versus the logarithm (base 10) of city size for A) 2016 and B) 1996. Visual inspection immediately suggests no correlations of the variance in the $\xi(t)$ with city population size. Dividing the data into quartiles (delimited by vertical red dashed lines) allows us to estimate the standard deviation of each subsection of the data (shown as red dots). We observe very small variations of the estimated quartile standard deviations supporting the standard hypothesis that the data are homoscedastic and that OLS estimation in log-transformed variables (scaling analysis) is justified. Estimating a slope on the logarithmic population size for these four points result in a value of -0.0064 with 95% CI [-0.0157, 0.0029] with a low $R^2 = 0.23$. Black lines shows the zero mean $\langle \xi \rangle$ (solid) and the average standard deviation across quartiles (dashed). For all years, estimated mean $|\langle \xi(t) \rangle| < 10^{-12}$, while the $R^2 < 10^{-3}$ for the correlation between $\xi_i$ and $\ln N_i$.

versus city size, $\ln N$. This behavior, quantified in Figure S1 for 1996 and 2016, is typical of all other years. Thus, wage residuals of scaling analysis are homoscedastic, justifying the OLS fit in log-transformed variables, Figs 2A-B.

These tests allows us to further motivate the treatment of urban indicators available over time, through scaling analysis as discussed in the section "Scaling relations and statistical deviations". Scaling analysis (together with theory) produces quantities that justify the estimation of exponents by OLS and construct their well-defined cross-sectional variations $\xi_i(t)$ at each time, in the standard sense discussed here. However, longitudinally (over time) all variables

- urban indicators, prefactors, populations and residuals – show very strong correlations from year to year, Figure 2D. The procedure of the section "Scaling relations and statistical deviations" is aimed at removing systemic temporal (due to system-wide growth of wages and of population) and city-population size (scaling) dependences, and leave us only with a de-trended behavior of the residuals, which can then be studied on their own right. Without this procedure, inter-temporal fits of urban indicators will typically be affected by strong mutual temporal correlations.

Finally, the point of the paper is to systematically analyze the possibility that residuals' fluctuations do show city size dependence. The main text shows how these dependences may naturally arise out of the group aggregation of microscopic dynamics and determines their consequences for the exponents, which in general do *run* with city size and time.

The data for wages in US Metropolitan Statistical Areas are shown here to essentially not produce any observable city size or temporal dependence in $\beta$ because i) the magnitude of the volatilities (which control any city size dependence) relative to the means is very small, and ii) they show an even smaller, very noisy city size dependence. The discussion section emphasizes that these features of wages do not necessarily apply to other urban indicators, some of which are known to be more volatile (e.g. murders or patents) or to other urban systems. The manuscript proposes a quantitative formalism across scales in which all these situations can be analyzed and thus identifies a future agenda for empirical studies with appropriately disaggregated long timeseries data.

## 2 Correspondence between PID Control and Driven Stochastic Oscillator

The stochastic dynamics of the PID controller error was shown in the main text to be

$$k_D \frac{d^2 \epsilon_r}{dt^2} + (k_P + 1)\frac{d\epsilon_r}{dt} + k_I \epsilon_r = \frac{dv}{dt}. \tag{S1}$$

Here we show that this is a familiar model of stochastic dynamics, the equation for a damped harmonic oscillator driven by the force $F = \frac{dv}{dt}$. Compare to the canonical equation for a harmonic oscillator with $\epsilon_r$ assuming the role of the displacement coordinate,

$$\frac{d^2\epsilon_r}{dt^2} + 2\zeta\omega\frac{d\epsilon_r}{dt} + \omega^2\epsilon_r = F/m, \qquad (S2)$$

with the correspondences: $m = k_D$, $\omega = \sqrt{k_I/k_D}$ and $\zeta = \frac{k_P+1}{2\sqrt{k_D k_I}}$, where $m$ is the mass, $\omega$ the oscillator's natural frequency and $\zeta$ the (dimensionless) damping ratio.

Because this is a linear ordinary differential equation, it is straightforward to integrate it and obtain a solution for $\epsilon_r(t)$ given the dynamics of the external field, $v(t)$. In particular, we can make a general statement that, if $v(t)$ and $\epsilon_r$ are constants in time, then they must be zero. Note that this only follows when $k_I$ is non-zero, i.e. when integral control is implemented. In particular, the derivative component of the PID controller is not essential. Setting it to zero, $k_D = 0$, makes the dynamics of the error simpler, corresponding to the well-known case of an over-damped driven oscillator.

We are particularly interested in the situation when the force resulting from income fluctuations is stochastic and cannot be known in advance in detail, only statistically. Other cases are straightforward and treated at length in several excellent sources [25] but are less general and will not be explored further in this paper. Suppose then that $\frac{dv}{dt}$ is normal white noise with variance $\Omega^2$. Then Eq. (S2) describes an Ornstein-Uhlenbeck process (an overdamped driven oscillator) [25], which is

$$d\epsilon_r = -\frac{M}{2}\epsilon_r dt + \Omega dW(t), \qquad (S3)$$

with $M = \omega/\zeta = k_I/[2(k_P+1)]$ and $dW(t)$ is normal white noise with unit variance. It is well known that the exact solution for the probability of $\epsilon_r$ at time $t$, given $\epsilon_0$ at the initial time, is

$$P[\epsilon_r, t|\epsilon_0] = \sqrt{\frac{M}{2\pi\Omega^2(1 - e^{-Mt})}} e^{-\frac{M}{2\Omega^2}\left[\frac{(\epsilon_r - \epsilon_0 e^{-M/2 \ t})^2}{1 - e^{-Mt}}\right]} \rightarrow \sqrt{\frac{M}{2\pi\Omega^2}} e^{-\frac{M}{2\Omega^2}\epsilon_r^2}, \qquad (S4)$$

where the arrow indicates the limiting behavior for times $t >> M^{-1} = \frac{\zeta}{\omega} = 2\frac{k_P+1}{k_I}$. This emphasizes again the role of integral control set by $k_I$. In general we want to make $k_I > 1$ and let the other parameters be small or zero, emphasizing the essential role of smoothing expenses over time in order to decrease stochastic income fluctuations.

## 3 Asymptotic Statistics of $r_G$

Here, we clarify and provide a numerical example illustrating the conditions under which population aggregate resources, $r_G$, approach limiting statistics given by a lognormal distribution.

First, it is critical to stress that none of the indicators for resources, income or costs needs to obey any particular form of statistics as a starting point. Their empirical observation over short times, especially for flow quantities, is likely to not reveal any simple general statistics. The argument of the paper is that agents –by necessity– will create strong correlations between incomes and costs, so that stochastic growth rates for resources will show a finite *small* variance, a process described in the text in terms of a simple control theoretical protocol that balances a resource account over a long period of time. Thus, consumption will be evened out to match incomes, and hopefully provide a surplus, but only over sufficiently long times. Then, fluctuations in the growth rate for resource accumulation will average out over time and, as a consequence of the central limit theorem, converge to a Gaussian variable. This predicts that $r_i(t)$ approaches, in the same limit and for the same formal reasons, a lognormal distribution.

Second, an analogous but different set of arguments applies to the asymptotic statistics of the population averaged resources, $r_G$. This variable is a cross sectional average - the mean of a set of random exponentials– each for a different agent. Such a mean will have different *asymptotic* statistical behavior depending on various limits taken on both group size $G$, and time, $t$. For example, it is well known that by taking joint limits where $G$ and $t$ go to infinity together, $r_G$ typically approaches standard behavior predicted by the law of large numbers and/or the central

limit theorem, in some cases with shifted amplitudes, see [29]. Mathematically, this is entirely consistent with the results of this manuscript but requires taking additional *infinite* time and group size limits.

In contrast, to describe the actual statistics of cities and urban systems we must *not* take either limit and work instead with *finite* populations and times. For *intermediate* times $t \sim t_*$ - not instantaneous, but also not infinite - and finite group sizes, the statistics of $r_G$ will then depend on both $t$ and $G$, as derived in Eqs. (12-14). The important point of these expressions is that they make it clear when the statistics of $r_G$ will approach a relatively simple lognormal limit, as well as conditions when such convergence may fail.

Specifically, for the lognormal limit to arise requires the existence of well-defined temporal means for $\eta'_G, \sigma'^2_G$, and most critically errors, $\epsilon'_G$, that obey the conditions of the standard central limit theorem. The claim of this manuscript is that such a regime exists under specific conditions, which arise naturally for a set of interacting agents dynamically balancing their budgets, as people must do in cities. This convergence may be aided in practice by collective resource flow management, for example by firms and above all by local and national governments, but these additional levels of control are not essential for our argument.

Note, however, that the convergence of $r_G$ to a lognormal described in the section "Population Dynamics and Emerging Inequality" is not unconditional or general outside our context. In terms of the behavior of the error, $\epsilon'_G$, both exceptional individuals and correlations between growth rates and resource magnitudes have an influence in the non-trivial behavior of the group averages, the latter being the explicit result of the non-linear (multiplicative) character of the errors relative to resources. Both of these types of effect raise questions that must be settled empirically. Nevertheless, we can supply some consistency arguments: To the first issue, it is highly unlikely that exceptional individuals can sustain over time errors so large in their growth rates so as to tilt the entire group averages given our discussion of the budget condition in mul-

tiplicative growth. To the second issue, the covariance between resources and growth rates is a real concern and naturally leads to microscopic distributions of wealth in a population that are not lognormal, especially at the richer end of the scale because in reality rich individuals often beget larger growth rates. However, it is a sufficient (much weaker) condition for the group mean $r_G$ to approach a lognormal distribution at long times that the covariance term in Eq. (13) be finite (non-zero) and a constant in $r_G$. Whether it is or not, is an empirical question calling for data on individuals' budgets over time that should be investigated in future work with more disaggregated evidence.

To explicitly demonstrate that $r_G$ approaches a lognormal variable under these conditions, we show in Figure S2 an example using stochastic simulations of an ensemble of 3000 geometric random walkers, as in Figure 3B (main text). In this case, the covariance term is zero and the means for the growth rate and volatility are standard. For statistically independent agents, the standard deviation $\sigma'_G$ becomes very small in large groups (decaying with $1/\sqrt{G}$). This suggests that, in cities, agents experience instead many shocks in common, as may be expected from shared (financial, real estate) markets and labor opportunities mediated by an integrated economy at the metropolitan level. This is another way of appreciating the importance of functional definitions of cities (Metropolitan Statistical Areas), as integrated labor markets.

# 4 Derivation of Equations of motion for prefactors and residuals

We start by deriving the equation of motion for the residuals, $\xi_i^r$. We take total resources in the city as a scaling relation, $R_i(N_i, t) = R_0(t)N_i(t)^\beta e^{\xi_i^r(t)}$. Then,

$$\frac{d\xi_i^r}{dt} = \frac{d\ln R_i}{dt} - \frac{d\ln R_0 N_i^\beta}{dt} = \frac{d\ln R_i}{dt} - \frac{d\langle\ln R\rangle}{dt} - \beta\frac{d\ln N_i}{dt} + \beta\frac{d\langle\ln N\rangle}{dt}, \quad (S5)$$
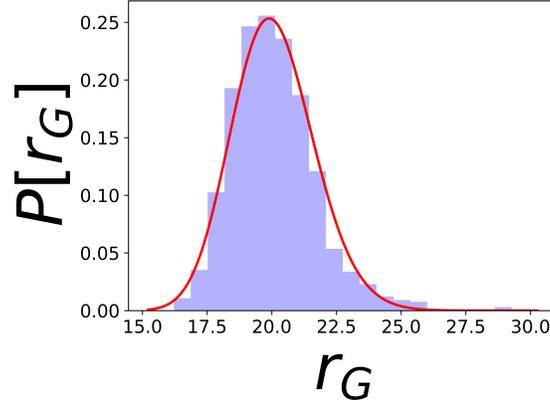
Figure S2: **The statistics of the mean growth rate for a group $r_G$, such as a city, is lognormal at sufficiently long times**. When the growth rates for an ensemble of agents can be averaged to produce an aggregate rate that obeys the central limit theorem conditions Eqs. (12-14), then the group average $r_G$ approaches a lognormal variable (red line) for long times, with parameters dependent on group characteristics. Here, we used the same set of geometric random walkers as in Fig. 3B, and produced the mean $r_G$ for a group of size 3000, over 1000 independent realizations.

which implies that the intensive resource growth rate (net of growth due to agglomeration effects) is

$$\frac{d \ln R_0}{dt} = \frac{d \langle \ln R \rangle}{dt} - \beta \frac{d \langle \ln N \rangle}{dt}. \tag{S6}$$

This is a function of the centers' dynamics only. This quantity, moving the scaling lines up or down in Figure 2A, is a measure of endogenous change net of agglomeration effects due to population. Next, let us introduce explicit growth dynamics $\frac{d \ln R_i}{dt} = \gamma_i + \epsilon_i$,

$$\frac{d \xi_i^r}{dt} = \gamma_i + \epsilon_i - \frac{d \ln R_0 N_i^\beta}{dt}. \tag{S7}$$

Note that the presence of $\gamma_i$ recognizes that we are dealing with stochastic calculus. Now take the average over cities to get

$$\gamma = \frac{d \langle \ln R_0 N_i^\beta \rangle}{dt} - \epsilon = \frac{d \langle \ln R_i \rangle}{dt} - \epsilon. \tag{S8}$$

Substituting (S8) into (S7) results in

$$\frac{d\xi_i^r}{dt} = (\gamma_i - \gamma) + (\epsilon_i - \epsilon) - \beta \frac{d}{dt} (\ln N_i - \langle \ln N \rangle) \tag{S9}$$

$$= \gamma_i - \gamma - \beta (\gamma_{N_i} - \gamma_N) + (\epsilon_i - \epsilon),$$

where $\gamma_{N_i} = \frac{d}{dt} \ln N_i$ and $\gamma_N = \frac{d}{dt} \langle \ln N \rangle$.

Finally, we can integrate this relation under the same conditions of the main text to obtain

$$\xi_i^r(t) - \xi_i^r(0) = [(\gamma_i - \gamma) - \beta (\gamma_{N_i} - \gamma_N)] t + (\sigma_i - \sigma) W(t), \tag{S10}$$

which is just Brownian motion (a random walk) with drift.

We can use this result to show that

$$\Delta_i^r = \frac{1}{t} (\xi_i^r(t) - \xi_i^r(0)) - [(\gamma_i - \gamma) - \beta (\gamma_{N_i} - \gamma_N)], \tag{S11}$$

is ergodic in the sense discussed above, since the stochastic expectation value of $\langle (\Delta^r)^2 \rangle \rightarrow \frac{(\sigma_i - \sigma)^2}{t} \rightarrow 0$, for long times. Note, however that the ensemble deviations $\xi_i^r$ are not necessarily ergodic in the same strict sense of the $\epsilon_i$ and that the difference depends principally on the rates of resource and population growth conforming with their ensemble averages over a given time period. i.e. on the dynamics of inequality between cities.

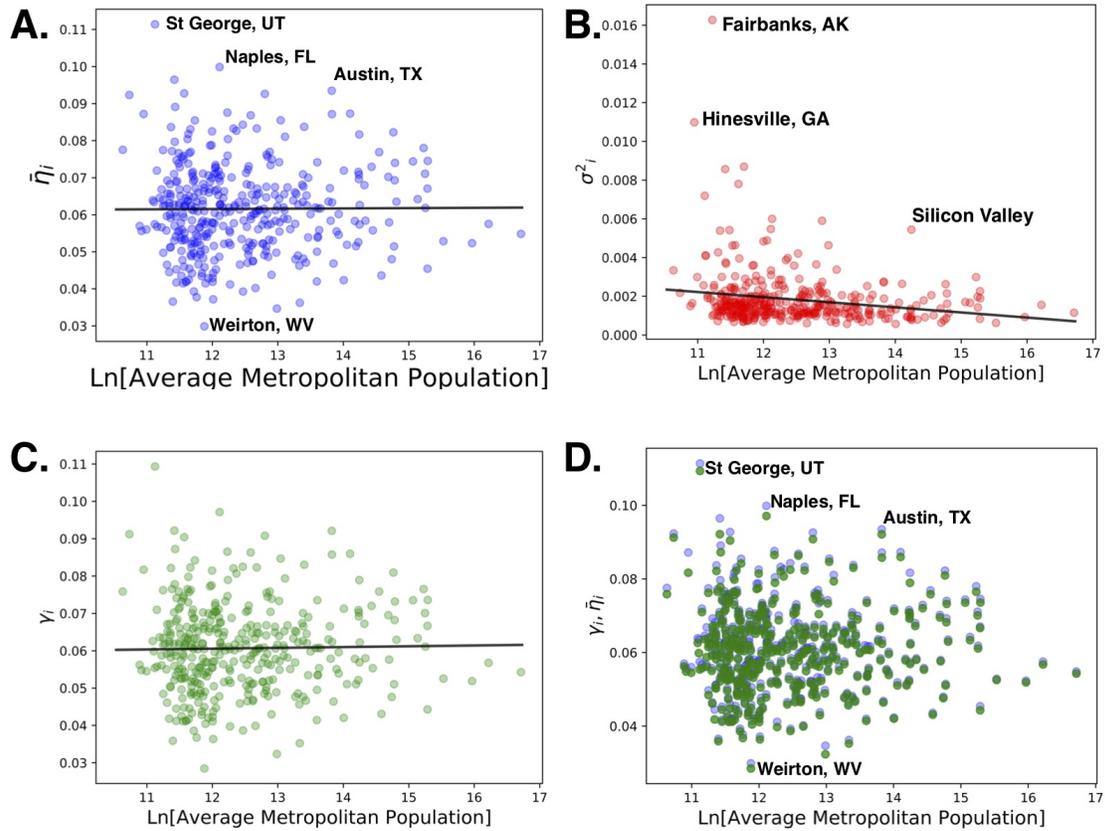# 5 Measured growth parameters for US MSAs and dynamics of deviations

Figure S3: **Measured growth parameters for US MSAs in the period 1969-2015.** A. The average growth rate $\bar{\eta}_i$ is statistically independent of city size (gradient= 0.00008 , 95 % confidence interval (CI) = [-0.0011,0.0013]; intercept= 0.0606, 95% CI [0.0459,0.0752]; $R^2 = 2.7e^{-9}$). B. The growth rate square volatilities $\sigma_i^2$ display a very small negative dependence on city size (Gradient= -0.00026, 95% CI [ -0.00040,-0.00013], intercept= 0.00512, 95% CI =[0.00341,0.00683], $R^2 = 0.00142$), due primarily to a few high-volatility small cities indicated. C. The effective growth rate $\gamma_i = \bar{\eta}_i - \frac{\sigma_i^2}{2}$ shows no significant population size dependence (gradient= 0.00021, 95 % CI = [-0.00093,0.00136] intercept=0.05799, 95% CI = [ 0.04365,0.07233], $R^2 = 1.34e - 07$). D. The effective growth rate and mean growth rate for each city, blue and green dots, respectively. The contribution of the volatilities to the total effective growth rates is very small in most cases, with a few outstanding examples noted.
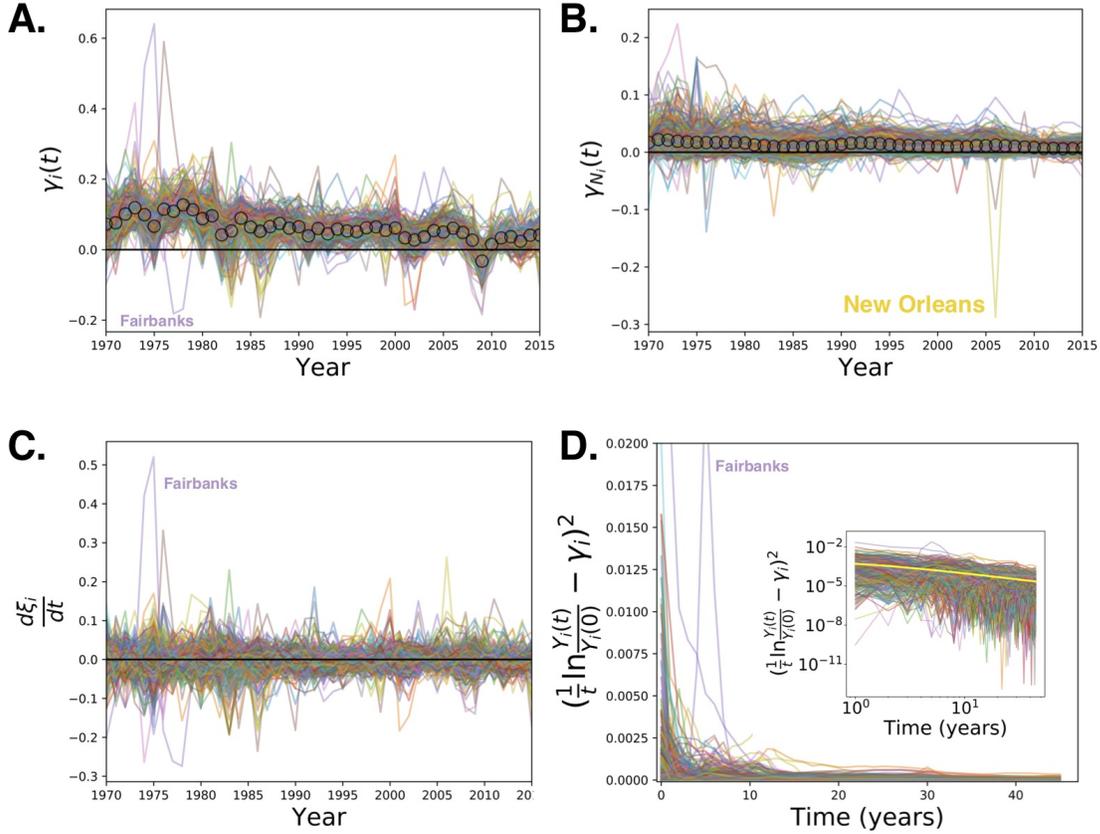
Figure S4: **Measured growth rates for US MSAs and the dynamics of the deviations** $\xi_i(t)$.
A. The effective growth rate $\gamma_i(t)$ for wages measured each year, for every US MSA (different colors). Open circles show the averages over all cities, $\gamma(t)$. B. The effective growth rate $\gamma_{N_i}(t)$ for population measured each year for every city (different colors). As in panel A, open circles show the averages over all cities, $\gamma_N(t)$. The salient negative growth of New Orleans shows the temporary effects of Hurricane Katrina. C. The dynamics of deviations $\xi_i$, according to Eq. (S9). Observe how there is now no mean growth rate so that the dynamics becomes purely diffusive (no drift). D. Measured ergodicity of wages' growth: At long times the effect of fluctuations averages out and growth in $Y_i(t)$ is as expected from the pure effect of exponential growth under a constant $\gamma_i$. The inset shows that the decay of the fluctuations is given on average by the pure Brownian diffusive expectation $\sim t^{-1}$, though in detail different cities show specific dynamics often driven by non-random events.

# 6 Summary of Mathematical Notation

$\langle x \rangle = \frac{1}{N_c} \sum_{i=1}^{N_c} x_i$; average over cities in an urban system with $N_c$ cities.

$x_G = \frac{1}{G} \sum_{j=1}^{G} x_i$; average over individuals in group of size $G$.

$x_{N_i} = \frac{1}{N_i} \sum_{j=1}^{N_i} x_i$; average over individuals in city of size $N_i$.

$\bar{x} = \frac{1}{t} \sum_{l=1}^{t} x(t_l)$ average over time in interval $t_l \in [0, t]$.

Table S1: Variables and their units. Breaks separate different sections in the main text.

| symbol | notation | units |
|---|---|---|
| $Y_i$ | total income in city $i$ | resources/time |
| $Y_0$ | $Y$ scaling intercept $i$ | resources/time |
| $N_i$ | population in city $i$ | number of people |
| $R_i$ | total resources in city $i$ | resources |
| $\beta$ | scaling exponent | 1 |
| $\xi_i$ | $Y$ scaling residuals for city $i$ | 1 |
| $\xi_i^r$ | resource scaling residuals for city $i$ | 1 |
| $r$ | individual accumulated resources | resources |
| $y, c$ | individual income, costs | resources/time |
| $\eta_r$ | resources' stochastic growth rate | 1/time |
| $\bar{\eta}_r$ | $\eta_t(t)$ temporal average | 1/time |
| $\epsilon_r$ | growth rate fluctuations | 1/time |
| $\sigma_r^2$ | volatility of resources' growth rate | 1/time |
| $\gamma_r$ | resources' effective growth rate | 1/time |
| $\Theta(t)$ | sum of $\epsilon_r$ over period $t$ | 1/time |
| $t_*$ | critical growth time | time |
| $b$ | $y/r$, income to resources ratio | 1/time |
| $a$ | $c/r$, costs to resources ratio | 1/time |
| $\bar{b}$ | average $b(t)$ over period $t$ | 1/time |
| $\bar{a}$ | average $a(t)$ over period $t$ | 1/time |
| $v$ | fluctuations in $b(t)$ | 1/time |
| $u$ | control variable, fluctuations in $a(t)$ | 1/time |
| $k_P$ | proportional control parameter $t$ | 1 |
| $k_I$ | Integral control parameter $t$ | 1/time |
| $k_D$ | Derivative control parameter $t$ | time |
| $M$ | rate fluctuations "mass" | 1/time |
| $\Omega^2$ | variance of fluctuations' driving force | 1 |
| $W(t)$ | Wiener process over time $t$ | 1/time |
| $G$ | group size | number of individuals |
| $x_G(t)$ | average quantity $x$ in $G$ at time $t$ | resources |
| $\rho_{jk}$ | correlation matrix for rate within $G$ | 1 |
| $\eta_G'$ | effective average growth rate within $G$ | 1/time |
| $\gamma_i$ | effective resources growth rate city $i$ | 1/time |
| $\gamma$ | average $\gamma_i$ across cities | 1/time |
| $\gamma_{N_i}$ | average population growth rate city $i$ | 1/time |
| $\gamma_N$ | average $\gamma_{N_i}$ across cities | 1/time |
| $\epsilon_i$ | fluctuation in population average growth rate city $i$ | 1/time |
| $\epsilon$ | average $\epsilon_i$ across cities $i$ | 1/time |
| $B(\ln N)$ | scale running function for $\gamma$ | 1/time |
| $\Delta^{r,y}$ | ergodic invariant for $r$ or $y$ | 1/time |