Research

Cite this article: Bettencourt LMA, Yang VC, Lobo J, Kempes CP, Rybski D, Hamilton MJ. 2020 The interpretation of urban scaling analysis in time. *J. R. Soc. Interface* **17**: 20190846.
<http://dx.doi.org/10.1098/rsif.2019.0846>

Received: 11 December 2019

Accepted: 10 January 2020

Subject Category:

Life Sciences—Mathematics interface

Subject Areas:

biocomplexity, environmental science

Keywords:

urban scaling, statistics, cities, congestion, regression, growth

Author for correspondence:

Vicky Chuqiao Yang

e-mail: vcy@santafe.edu

The interpretation of urban scaling analysis in time

Luís M. A. Bettencourt^{1,2}, Vicky Chuqiao Yang², José Lobo³, Christopher P. Kempes², Diego Rybski⁴ and Marcus J. Hamilton^{2,5}

¹Mansueto Institute for Urban Innovation, Department of Ecology and Evolution and Department of Sociology, University of Chicago, Chicago, IL 60617, USA

²Santa Fe institute, 1399 Hyde Park Road, Santa Fe, NM 87501, USA

³School of Sustainability, Arizona State University, 800 South Cady Mall, Tempe, AZ 85281, USA

⁴Potsdam Institute for Climate Impact Research - PIK, Member of Leibniz Association, PO Box 601203, Potsdam 14412, Germany

⁵Department of Anthropology, University of Texas at San Antonio, San Antonio, TX 78202, USA

VCY, 0000-0001-6111-3699; CPK, 0000-0002-1622-9761; MJH, 0000-0003-1754-6969

Scaling is a general analytical framework used by many disciplines—from physics to biology and the social sciences—to characterize how population-averaged properties of a collective vary with its size. The observation of scale invariance over some range identifies general system types, be they ideal gases, ecosystems or cities. The use of scaling in the analysis of cities quantifies many of their arguably fundamental general characteristics, especially their capacity to create interrelated economies of scale in infrastructure and increasing returns to scale in socio-economic activities. However, the measurement of these effects, and the relationship of observable parameters to theory, hinge on how scaling analysis is used empirically. Here, we show how two equivalent approaches to urban scaling—cross-sectional and temporal—lead to the measurement of different mixtures of the same fundamental parameters describing pure scale and pure temporal phenomena. Specifically, temporal exponents are sensitive to the intensive growth of urban quantities and to circumstances when population growth vanishes, leading to instabilities and infinite divergences. These spurious effects are avoided in cross-sectional scaling, which is more common and closer to theory in terms of quantitative testable expectations for its parameters.

1. Introduction

Scaling is a general analytical framework to model how aggregate properties of systems depend on measures of their size [1]. The use of scaling analysis is pervasive as a method across disciplines—physical, biological and social—starting with equations of state for gases and liquids [2], to biological metabolism, to populations in ecology and anthropology [3,4] and to the properties of firms and cities [5,6]. In biology, and when applied to morphological quantities in other systems such as cities, scaling is often referred to as *allometric*, but in general scaling analysis also applies to other quantities such as social networks, energy dissipation or congestion.

When scale-invariance relative to measures of size (often expressed as power laws, including linearity) emerges as an average characteristic property of any of these systems, we speak of general system types, existing across scales, such as gases, organisms, ecosystems or cities. All these system types manifest *different* scaling relations, accounting for their specific nature; while sharing the common quantitative property that each type displays some average scaling characteristic. The empirical identification of scaling relations for specific system types is critical as the basis for theory in every domain, from thermodynamics (as equations of state [7]), to biological metabolism [8,9], to ecological biodiversity [10–13] or to network effects in human societies [14–17]. This criticality is because scaling relations quantify the average, macroscopic effects of a set of common microscopic

processes, be they laws of motion, interactions and conservation in liquids, gases, biological or social systems.

The application of scaling to cities has been particularly fertile recently as data in many urban systems throughout the world—past and present—have become increasingly available. Urban scaling analysis reveals how *general* nonlinear properties of how cities work, manifested as economies of scale, when certain quantities grow more slowly than city size (sublinearly) or as *increasing returns to scale*, when quantities grow faster than city size (superlinearly) [5]. Moreover, the scaling exponents characterizing these relationships, are fundamental to the study of urbanization, and have relatively simple interpretations. In addition, recent research suggests these general properties may characterize network effects in a diversity of human systems, from hunter–gatherer societies [4] to contemporary firms [18]. These discoveries provide an exciting context for developing a testable and predictive general science of cities. A number of theoretical models have been proposed as origins to superlinear scaling behaviours across cities [16,19–22], but only some of these predictions actually give detailed predictions for exponent values. The proposed mechanisms relate social interactions driven by functional complementarities (e.g. division of labour) taking place over built spaces and subject to spatial costs of movement and rents. These general dynamics lead to network effects, with some fractal signatures associated with the structure of networks and spatial densities. By accounting for general patterns in urban areas, scaling analysis also identified their exceptions as residuals, which account for the unique history and specificity of each place [23–26]. We note, however, that the importance of city scale (i.e. population size) as an important determinant of salient urban characteristics has long been noted in urban economics and geographers [27,28]. Also in this paper, we focus on using the population as the fundamental size measure of cities, though a number of previous studies have also examined space as a key variable, including fractal dimensions of cities [29,30].

The apparent simplicity of scaling analysis hinges on methodological and empirical choices which are not always made fully explicit (but see [5,16,26,31]) and have been questioned in the literature [32]. The main issue concerns how best to isolate conserved (or at least slowly changing) parameters characterizing general fundamental urban effects against a background of growth or decay (in population and urban quantities themselves) and of specific contextual effects in different places. These phenomena illustrate the central issue addressed in this paper. For open-ended, complex systems such as cities, time, as well as population size (and possibly other scales), plays a critical role because these systems are not in equilibrium. Instead, cities typically experience exponential growth rates of several per cent a year. Moreover, at larger timescales, the development of human systems, urban or otherwise, has been characterized by growth, not stasis, particularly over the last several thousand years of the Holocene, the current geological epoch. This means that *time*, along with population size, must generally be considered as another essential variable. Analytical methods that can separate the effects of time from those of population size are therefore particularly important. They can point to a number of independent urban phenomena: extensive growth or decay, which is change as a result of variations in scale due to agglomeration or network scaling effects [5,16,33–35]; and intensive growth or decay, change independent of scale, which is a fundamental target of economic theory [36–38].

Two different approaches to scaling analysis illustrate how time and population effects can be combined or separated in terms of parameters estimated and their connection to theory. The most common approach to urban scaling, here referred to as *cross-sectional*, holds time constant and compares the properties of all cities in an urban system as a function of their population. Cross-sectional scaling, therefore, isolates elasticities of urban properties relative to population size at fixed times. These elasticities or scaling exponents take similar values in many different circumstances [5,25,26,39–45] and are expected by theory to be stationary [16], up to higher-order statistical corrections [46]. *Temporal* scaling is the orthogonal approach to the cross-sectional procedure [32,47]. It takes the trajectory in time of a specific quantity for a particular city and attempts to fit it using a power-law function of its population scale, which itself is changing in time. In this way, temporal scaling analysis attempts to compare the growth in urban quantities with the growth of the population. Although there is an exact mathematical mapping between the parameters of these two approaches, as shown below, temporal scaling parameters can manifest instabilities and divergences, specifically if intensive growth is at play and population growth is small or vanishing, which is characteristic of many cities, especially in high-income countries.

Even though examples of urban scaling are now common in the scientific literature [5,33,39,48–51] and precursors have existed for decades in urban economics and economic geography (see the supplement in [16] for a review), recently questions have arisen in comparing these two approaches for urban scaling analysis. In particular, we are not aware of any contribution that shows explicitly (in terms of a mathematical expression) how estimated exponents following the two different methods map onto each other. Here, we present a short and synthetic description of these two approaches to urban scaling, show some empirical examples from the recent literature and provide the equivalence mapping and properties of the exponents obtained by the two procedures. We aim to clarify that the two types of scaling behaviour—cross-sectional and temporal—should not be expected to coincide as such a correspondence is not a general property of urban phenomena. Below, we derive the specific circumstances when this coincidence does happen and why.

2. Results

We now compare cross-sectional and temporal scaling analysis, to show their equivalence but also their separate meanings in terms of the parameters measured and their relation to general theoretical expectations about scaling of urban attributes. Before proceeding, we make a clarifying comment on methodology. The research design involving repeated observations of the same variables over a period, that is the combination of cross-sectional and time-series data, is quite common in economics, epidemiology, and psychology. But the analytical attention of *panel estimation* centres on controlling for the effects of omitted variables that differ across observational units or across time, and on controlling for heteroskedasticity within cross-sectional panels and time-series ensembles [52–54]. Here, we want to first highlight the difference between cross-sectionally and temporally estimated parameters (scaling coefficients), and then clarify the relationship between two

Figure 1. Cross-sectional scaling of extra congestion costs with population for US urban areas, for years between 1982 and 2014. (a) Log-log plot of extra congestion versus the population of cities with colours showing different years, 1982–2014, blue to orange. Yellow squares show the centres of each (x, y) -coordinate ($\langle \log Y \rangle$, $\langle \log N \rangle$) in each year, see text. The inset shows the time trajectory of the prefactor, $\log Y_0(t)$. (b) The variation of the scaling exponent, $\beta(t)$, in time. Points indicate best-fit estimates with bars showing the 95% confidence interval around this central value. Although extra congestion is not a fundamental quantity in urban scaling theory, congestion costs are typically predicted to scale with an exponent, $\beta \approx 7/6$ to which the estimate seems to converge over the last decade; it remains an empirical question for future studies if this trend will continue. This specific value of the scaling exponent has the important property that congestion costs, even as they increase *per capita* with city size, remain a fixed fraction of the city's gross domestic product (the aggregate value of its economy). (c) Centred scaling, where the time variation of the centres has been subtracted so that all data clouds are centred at $(0, 0)$ (yellow square). This procedure creates a single parameter estimation problem for the slope, $\beta(t)$. The inset shows the variation in time for the goodness of fit (R^2) for these cross-sectional scaling relations relative to theory. (d) The residuals, $\xi_i(t)$, for each city over time. We see that these quantities, with a few exceptions, are relatively stable in time and vary relatively little over timescales of decades as observed for other urban indicators elsewhere [23,46]. (Online version in colour.)

distinct ways in which scale can determine urban behaviours. By focusing attention on the effects of scale, we are intentionally omitting other variables which surely also influence urban characteristics; we do this in recognition of the importance of system size on system behaviour [55].

2.1. Cross-sectional urban scaling

The main technical issue for urban scaling analysis is that there are (at least) two *extensive* scales with different characters: population size and time. Extensive here refers to the property of a variable to account for the size of the system (not necessarily linearly). Cross-sectional urban scaling is the most common procedure for urban scaling analysis, as it identifies parameters that are averages over the system of cities and their deviations at a fixed time, see [5,26,31] and supplement in [16] for examples. This procedure can be repeated at consecutive times to derive the time dependence of the same parameters. In this approach, we write any extensive urban

quantity, Y , for a set of cities labelled $i = 1, \dots, N_c$ (N_c is the number of cities in the urban system), cross-sectionally

$$Y(N_i(t), t) = Y_0(t) N_i(t)^{\beta(t)} e^{\xi_i(t)}, \quad (2.1)$$

at the *same* time t , to estimate scaling exponents, $\beta(t)$, and prefactors $Y_0(t)$, see figure 1a. The prefactor subsumes the factors of size-independent effects, such as technology or institutional arrangements, which modulate how population size affects the baseline of attributes captured by the variable Y [56]. This analysis is then performed for every time period (usually consecutive years), $t = 1, \dots, T$, so that the time dependence of the population in each city, $N_i(t)$, and the prefactor, $Y_0(t)$, are tracked over time. Naturally, this also allows us to obtain the time dependence of the exponent $\beta(t)$, see figure 1b, and the deviations from scaling for each city, $\xi_i(t)$ (figure 1d).

In general, both population and the prefactor, $Y_0(t)$, are observed to vary exponentially in time, $Y_0(t) = Y_0(0) e^{\gamma(t)t}$, $N_i(t) = N_i(0) e^{\gamma_i(t)t}$, with the corresponding stochastic growth

rates, $\eta(t)$, $\gamma_i(t)$, typically taking values of a few per cent a year and fluctuating by a fraction of that amount. In modern urban systems, it is usually true that the temporal growth of quantities such as the size of the economy (e.g. gross domestic product (GDP)) or the built-up area of cities is faster than population growth. In ancient and pre-industrial urban systems, both rates were minimal (at most, fractions of a per cent per year) and the growth of the prefactors, Y_0 , could often be neglected, see [41] for an explicit example with rare longitudinal archaeological data.

The scaling exponent, $\beta(t)$, is often the quantity of greatest interest because it expresses the average (over all cities at the same time) nonlinear effects of population-scale on the characteristics of all cities. Technically, the scaling exponent is a rather familiar quantity: it is an average *elasticity* at fixed time; a derivative of the logarithm of a quantity relative to the logarithm of population at fixed time:

$$\beta(t) = \left. \frac{d \log Y(N_i, t)}{d \log N_i(t)} \right|_{t=\text{constant}}. \quad (2.2)$$

This follows from taking the logarithm of equation (2.1),

$$\log Y(N_i, t) = \log Y_0(t) + \beta(t) \log N_i(t) + \xi_i(t), \quad (2.3)$$

and then the derivative of both sides of the equal sign relative to $\log N_i(t)$. Since both the prefactor, $Y_0(t)$, and the variations, $\xi_i(t)$, are independent of the population scale, $N_i(t)$, the scaling exponent is given by the expression in equation (2.2). This independence of scale is very convenient, as we shall see below, because it means necessarily that these quantities must account for other urban dynamics, including intensive growth, that are by definition independent of population and of the dynamics specific to each city.

Given these simple properties, it is particularly convenient to further simplify the scaling relation by *centring the data* [26], which is done by defining the averages over all cities in the urban system

$$\begin{aligned} \langle \log Y \rangle(t) &= \frac{1}{N_c} \sum_{i=1}^{N_c} \log Y(N_i, t) \\ \text{and } \langle \log N \rangle(t) &= \frac{1}{N_c} \sum_{i=1}^{N_c} \log N_i(t). \end{aligned} \quad (2.4)$$

We can now take the average over cities in equation (2.3) to obtain

$$\langle \log Y \rangle(t) = \log Y_0(t) + \beta(t) \langle \log N \rangle(t), \quad (2.5)$$

where the variations $\xi_i(t)$ disappear because by definition, $\langle \xi \rangle(t) = 0$. This allows us to write the scaling relation in an even simpler form as

$$\Delta \log Y_i(t) = \beta(t) \Delta \log N_i(t), \quad (2.6)$$

which now depends on a *single parameter* $\beta(t)$, where $\Delta \log Y_i \equiv \log Y(N_i(t), t) - \langle \log Y \rangle(t)$ and $\Delta \log N_i(t) \equiv \log N_i(t) - \langle \log N \rangle(t)$ (figure 1c). We can also see that this expression gives us a direct way to estimate the exponent as

$$\begin{aligned} \hat{\beta}_i(t) &= \left. \frac{\Delta \log Y_i(t)}{\Delta \log N_i(t)} \right|_{t=\text{constant}} = \frac{\beta(t) \Delta \log N_i(t) + \xi_i(t)}{\Delta \log N_i(t)} \\ &= \beta(t) + \frac{\xi_i(t)}{\Delta \log N_i(t)}, \end{aligned} \quad (2.7)$$

where $\hat{\beta}_i(t)$ is an estimator of the exponent, $\beta(t)$. This is noisy

but the variations vanish because $\langle \xi \rangle = 0$, leading to $\langle \hat{\beta} \rangle(t) = \beta(t)$ (note that the variations ξ_i are uncorrelated to scale by construction). Given an estimate of $\beta(t)$, the prefactor $\log Y_0(t)$ can be computed immediately from equation (2.5) given also the averages over cities, equation (2.4).

To illustrate an example of this procedure on data, we analyse a dataset of extra costs due to vehicle congestion in US urban areas, collected by the Texas A&M Transportation Institute (see Material and methods). Road congestion is important for cities because it is a component of the transportation costs that must be balanced by incomes self-consistently [16,56]. The quantity estimated is a little peculiar itself, because it is not congestion *per se*, but an estimate of extra costs. It is also not about the entire transportation system of a city but just its cars and trucks. The urban definitions used in the dataset are approximations to US Metropolitan Statistical Areas, which are the standard definitions of functional cities in the USA, but in some cases there are some differences. Nevertheless, we think this dataset serves as an illustrative example for two reasons. First, the dataset has a long history (annually since 1982), maintains consistency, and is openly available. Second, this dataset provides a compelling illustration of varying scaling exponents over time and seemingly different results obtained by using cross-sectional and temporal scaling methods, as illustrated in [32].

2.2. Temporal scaling and its relationship with cross-sectional scaling

A number of studies have recently performed a different (but as we shall see, fundamentally equivalent) form of scaling analysis attributing changes *in time* for the quantities $Y(N_i(t), t)$ to changes in $N_i(t)$ [32,57]. We shall refer to this approach as *temporal scaling*. This approach starts with temporal averages for each city and then repeats the analysis for all cities, thus following the opposite order of operations in treating population size and time relative to the analysis above.

Specifically, temporal scaling analysis posits a scaling relation of the form

$$Y(N_i(t), t) = A_i N_i(t)^{a_i} e^{\chi_i(t)}, \quad (2.8)$$

where crucially the prefactor, A_i , is city-specific but *independent of time*. As shown below, this tends to make the scaling exponent a_i , which is also independent of time but specific for each city, vary strongly between different urban areas. Examples are shown in figure 2. The variations $\chi_i(t)$ no longer vanish in the average over cities but do vanish for each city under time averaging, as they are residuals of a temporal fit. Because both forms of scaling relations are exact (when the variations are included), we can map their parameters to each other. First, take the logarithm of equation (2.8) to obtain

$$\log Y(N_i(t), t) = \log A_i + a_i \log N_i(t) + \chi_i(t). \quad (2.9)$$

We can now use the analogous procedure as in the cross-sectional case by subtracting the averages,

$$\begin{aligned} \overline{\log Y_i} &= \frac{1}{T} \sum_{t=1}^T \log Y(N_i(t), t) \quad \text{and} \\ \overline{\log N_i} &= \frac{1}{T} \sum_{t=1}^T \log N_i(t), \end{aligned} \quad (2.10)$$

Figure 2. Temporal scaling relationship between extra traffic congestion costs and city population over time. Panels show a few examples with different characteristics. (a) New York City, which presents a typical temporal fit: observe how residuals are correlated in time typically larger for intermediate points (times) and lower for later ones. (b) Los Angeles, which shows a poor fit overall, especially because of the drop in excess congestion in later years, with an extremely large instantaneous negative slope. Fits like these show temporal structure in the residuals and would thereby be rejected on statistical grounds, meaning that the fits are not overall linear. (c) Pittsburgh, PA a city that over the time period for which the data are given here experienced very little population change (including periods of population loss). Together with increasing congestion, this results in a very large and negative exponent, $a_i = -32.6$. (d) Summary of data and exponents (slope of black lines) for all cities in the dataset. We see some consistency for a number of cities in terms of values of the exponent a_i , but also extreme variations, with the most extreme noted. Post-industrial cities, such as Pittsburgh, Cleveland, Providence, Detroit and others typically show very large exponents (in absolute value) because of very small or negative population growth over the time period, see text. (Online version in colour.)

which are now *temporal*, denoted by an overbar. We centre the data, in analogy with the procedure in the previous section, to get a one-parameter scaling relation, depending only on the exponent a_i

$$\log Y_i(N_i(t), t) - \overline{\log Y_i} = a_i(\log N_i - \overline{\log N_i}), \quad (2.11)$$

with the exponent (slope)

$$a_i = \frac{\log Y_i(N_i(t), t) - \overline{\log Y_i}}{\log N_i(t) - \overline{\log N_i}}, \quad (2.12)$$

which can be evaluated on the average *over time*, since $\overline{\chi_i} = 0$. Substituting the cross-sectional scaling relation of equation (2.3), we obtain a suggestive form expressing the meaning of the temporal exponent as

$$a_i = \overline{\beta} + \frac{\log Y_0(t) - \overline{\log Y_0}}{\log N_i(t) - \overline{\log N_i}} + (\beta(t) - \overline{\beta}) \frac{\log N_i(t)}{\log N_i(t) - \overline{\log N_i}} + \frac{\xi_i(t) - \overline{\xi_i}}{\log N_i(t) - \overline{\log N_i}}, \quad (2.13)$$

where $\overline{\beta} = (1/T) \int_0^T \beta(t) dt$, is the temporal average of $\beta(t)$. Deriving equation (2.13) uses the fact that β and $\log N_i$ are

uncorrelated by construct: $\overline{\beta \log N_i} = \overline{\beta} \overline{\log N_i}$. Finally, it is useful to write some of these quantities in terms of temporal growth rates. For example,

$$\begin{aligned} \log Y_0(t) &= \log Y_0(0) + \eta(t)t \rightarrow \overline{\log Y_0} \\ &= \log Y_0(0) + \overline{\eta} \bar{t}, \end{aligned} \quad (2.14)$$

where $\eta(t) = (1/t) \int_0^t dt' \eta(t')$, is the partial average growth rate and $\overline{\eta} = (1/T) \int_0^T dt' \eta(t')$ is the average over the entire time period under consideration; the time $\bar{t}$ is the midpoint in the time period. We can analogously define

$$\begin{aligned} \log N_i(t) &= \log N_i(0) + \gamma_i(t)t \\ \rightarrow \overline{\log N_i} &= \log N_i(0) + \overline{\gamma_i} \bar{t}. \end{aligned} \quad (2.15)$$

Finally, we can write the temporal exponent as

$$\begin{aligned} a_i &= \overline{\beta} + \frac{(\eta(t) - \overline{\eta})t + \overline{\eta}(t - \bar{t})}{(\gamma_i(t) - \overline{\gamma_i})t + \overline{\gamma_i}(t - \bar{t})} + (\beta(t) - \overline{\beta}) \\ &\quad - \overline{\beta} \frac{\log N_i(t)}{\log N_i(t) - \overline{\log N_i}} + \frac{\xi_i(t) - \overline{\xi_i}}{\log N_i(t) - \overline{\log N_i}}. \end{aligned} \quad (2.16)$$

From the perspective of this expression, we can now clearly recognize a number of different contributions to the value

of the temporal exponent, appreciate some of its particular cases, and specify the precise circumstances when a_i coincides with the cross-sectional scaling exponent, β .

Specifically consider the following sequence of simplifications and correspondingly more particular situations, which may correspond to different scenarios of urban development:

Scaling residuals are time-independent or vanishingly small ($\xi_i(t) = \bar{\xi}_i$):

$$a_i \rightarrow \bar{\beta} + \frac{\bar{\eta} + (\eta(t) - \bar{\eta})t/(t - \bar{t})}{\bar{\gamma}_i + (\gamma_i(t) - \bar{\gamma}_i)t/(t - \bar{t})} + (\beta(t) - \bar{\beta}) \frac{\log N_i(t)}{\log N_i(t) - \log \bar{N}_i}. \quad (2.17)$$

Note how the time dependence of the cross-sectional scaling exponent, $\beta(t)$, induces a population scale dependence of a_i , a result derived by other means in [46].

The cross-sectional exponent is time-independent ($\beta(t) = \beta$):

$$a_i \rightarrow \beta + \frac{\bar{\eta} + (\eta(t) - \bar{\eta})t/(t - \bar{t})}{\bar{\gamma}_i + (\gamma_i(t) - \bar{\gamma}_i)t/(t - \bar{t})}. \quad (2.18)$$

Note that this condition is in addition to the condition above.

The growth rates for the scaling prefactor and city population are time-independent ($\eta(t) = \eta$, $\gamma_i(t) = \gamma_i$):

$$a_i \rightarrow \beta + \frac{\eta}{\gamma_i}. \quad (2.19)$$

The growth rate of the prefactor (intensive growth) is zero, while population growth remains non-vanishing ($\eta = 0$, $\gamma_i \neq 0$):

$$a_i \rightarrow \beta. \quad (2.20)$$

Each one of these simplifications may apply (and be assumed in modelling) separately from the others, of course. The point is that the two exponents, cross-sectional and temporal, only coincide when three conditions are met: there is no intensive growth of the quantity of interest ($\eta = 0$), the scaling exponent β is time-independent, and the scaling residuals $\xi_i(t)$ are either vanishingly small or time-independent. Of these three conditions, the first is by far the most important (figure 3).

In appendix A, we also present a simplified derivation for (2.19) and (2.20) under the special case of constant growth rates and no scaling deviations.

Although traffic congestion exhibits variable scaling exponents over time, both exponents and residuals for other salient urban properties have been observed to be quite stable [5,16,23], with the latter expected to be stationary on theoretical grounds [16] and are observed to typically change slowly on the timescale of decades [23]. For most modern cities—especially in high-income nations—intensive growth rates, such as for the economy as a whole, are several times larger than population growth rates, leading to potentially large deviations of a_i relative to β . This deviation in the two exponents is especially large when population growth rates become vanishingly small (the average population growth rate for US cities is now smaller than 1% per year, and it is smaller in Europe and high-income nations in Asia). As equation (2.19) shows, these small growth rates can lead to unstable exponent a_i , which can diverge to infinity. We see this effect

Figure 3. Estimates of the temporal exponent via equation (2.13) compared to direct ordinary least squares minimization. Blue points show estimates for the exponent a_i from equation (2.13) (solid black line is the one-to-one diagonal). The estimator provides the correct result on average but is sensitive to points where $\log N_i(t) - \log \bar{N}_i \approx 0$, which are more common for large estimated exponents. Some of the most extreme points, see figure 2d, are not shown because they would dominate the figure and render other variations hard to see (a complete version of the figure is provided in appendix C). The red points show the same estimator, neglecting the contribution of the residuals ξ_i (last term in equation (2.13)). We observe that this term introduces some variations, but its absence remains consistent with the OLS estimator. Green points show the estimator including only the first and third terms, while black the first and second terms of equation (2.13): we observe that the contributions from the second and third terms introduce large fluctuations in the estimate that partially cancel each other out, as can be easily observed via the approximate symmetry about the solid line. (Online version in colour.)

over the cities in the USA, which experienced very small or negative population growth over the last few decades, especially deindustrializing urban areas such as Pittsburgh, Cleveland and Detroit. This parameter instability is not, in general, a desirable property of any observable quantity. As equations (2.13) and (2.16) show, this unstable behaviour is *not fundamental*: it is the result of the misallocation of an endogenous intensive growth rate to scale effects of a city that is not growing, thus leading to inconsistency and associated spurious divergence.

3. Discussion

Cross-sectional and temporal analyses are among the most fundamental of methodological strategies for investigating variables capturing systemic behaviour. The former animates comparative analysis, while the latter renders dynamics visible. Cities are complex systems, demonstrating at once nonlinear effects of scale, scale-independent temporal changes, and other demographic and local factors. Cross-sectionality and temporality are both necessary perspectives to understand urban systems.

A clear theoretical understanding of the general properties of cities is necessary to make sense of diverse applications of scaling to urban data. Scaling analysis asks how scale affects systemic performance, and how these effects have cross-sectional and temporal dimensions. Whether scaling analysis is performed cross-sectionally or temporally thus matters as this decision leads to different measured parameters with

different meanings. Here, we have demonstrated that the way scaling is most typically done, cross-sectionally and regularly spaced over time, is equivalent to temporal scaling analysis, where changes in an urban quantity for a specific city over time is compared to temporal changes in that city's population change.

The resulting scaling exponents coincide in specific simplifying limits: zero intensive growth rate of the urban property [36,58], time-independent scaling exponents, and small or time-independent cross-sectional residuals. The scaling exponents, β , have a specific meaning in urban scaling theory in terms of the magnitude of network effects realized on average in cities. This is proportional to the relevant network's average degree, which has a theoretical upper bound at $\beta \leq \beta_{\max} = 2$, for a fully connected network (this limit is known as Metcalfe's law) [16]. However, if intensive change is attributed to population variation when in effect there is no population growth, the temporal exponents diverge to infinity with unstable signs depending on fluctuations, often resulting in cases when $|a_i| \gg \beta_{\max}$, as shown in figures 2–3. This behaviour illustrates how temporal scaling exponents acquire different meanings and do not directly reflect interaction networks in cities. Our result gives an explicit formalization of the relationship between the cross-sectional and temporal exponents, and is agnostic to the underlying mechanism of the scaling behaviour.

The deviation between temporal and cross-sectional scaling is a feature of only certain types of systems. Scaling relationships have been much studied in biological systems, typically using the mass of an organism to be the size measure. In mammals, the cross-sectional and temporal scaling exhibit the same exponent. For example, the scaling of total metabolic rate, $\beta = a_i = 3/4$, is found across species of different size and across developmental time for the same organism [59]. These findings suggest that changes in an animal's metabolism is fully due to the increase in size (extensive growth). For other organisms' features, the connection between temporal and cross-sectional scaling is more complicated. Biomass production rates in mammals exhibit identical relationships for temporal and cross-sectional scaling relationships, but only in the case of early organism growth after which temporal scaling breaks down [59,60] (see appendix B). Compared to the scaling of animals, scaling of cities has the notable feature of intensive changes, which is an important factor affecting the socio-economic outputs of cities. [5].

The divergences in the temporal exponents, a_i , are completely understandable analytically, as we have shown here. Whether one chooses one form of scaling analysis over another comes down to what questions are being posed about an urban system. It is typical for data on urban systems to have size variation spanning several orders of magnitude cross-sectionally and supporting significant statistics for scaling exponents, while few cities have grown orders of magnitude in recorded temporal data, leading to large fluctuations in estimated scaling exponents. The analysis performed here points to the much greater utility of estimating urban scaling parameters cross-sectionally: it produces exponents and residuals with greater temporal stability and the quality of fitting for scale-invariant relations is consistently better. Furthermore, the estimated scaling parameters have a more direct connection to theory, describing agglomeration and network effects in cities which have long been the subject of study in geography and economics.

4. Material and methods

Data on excess congestion costs and other interesting properties of automobile traffic in US urban areas are available from the Texas A&M Transportation Institute (<https://mobility.tamu.edu/umr/congestion-data/>), which updates the analysis and data annually as part of their report Urban Mobility Scorecard. The urban units reported in this dataset approximate a somewhat restricted (smaller) definition of Metropolitan Statistical Areas (LMAB 2019, private communication), which have been created and updated by the US Census Bureau systematically since the 1960s, as the fundamental urban socio-economic units nationwide.

Data accessibility. All data needed for our analysis are available from the website described in materials and methods.

Authors' contributions. L.M.A.B., V.C.Y., J.L., and D.R. contributed to mathematical derivations, L.M.A.B. performed data analysis, all authors contributed to writing the paper.

Competing interests. We declare we have no competing interests.

Funding. We thank the Arizona State University–Santa Fe Institute Center for Biosocial Complex Systems for meeting support. V.C.Y. acknowledges funding from the Santa Fe Institute Omidyar Fellowship. D.R. thanks the Alexander von Humboldt Foundation for financial support under the Feodor Lynen Fellowship.

Acknowledgements. We thank Scott Ortman's feedback on the manuscript. D.R. is grateful to the Leibniz Association (project IMPETUS) for financially supporting our research.

Appendix A. Simplified derivation for constant growth rates and no residual

The mathematical derivations in the main text are quite involved. To help demonstrate the main ideas of the manuscript, we provide a more straightforward derivation for a special case of the main text, with constant growth rates and no deviation from scaling. In this section, we will derive equations (2.19) and (2.20) of the main text with the simpler procedure.

In the special case where scaling residuals are negligible, the cross-sectional scaling is represented as

$$Y_i(t) = Y_0(t)N_i(t)^{\beta(t)}. \quad (\text{A } 1)$$

The temporal scaling exponent a_i can be represented as the slope of a city's trajectory on the log scale,

$$a_i = \frac{\log Y_i(t + \Delta t) - \log Y_i(t)}{\log N_i(t + \Delta t) - \log N_i(t)}, \quad (\text{A } 2)$$

where Δt is a small time interval. Plug equation (A 1) in equation (A 2), and Taylor expand $\beta(t + \Delta t)$ for small Δt : $\beta(t + \Delta t) = \beta(t) + \beta'(t)\Delta t + O(\Delta t^2)$. After some algebra, we have,

$$\begin{aligned} \log Y_0(t + \Delta t) - \log Y_0(t) &= [a_i - \beta(t)][\log N_i(t + \Delta t) \\ &\quad - \log N_i(t)] \\ &\quad - \beta'(t)\Delta t \log N_i(t + \Delta t) \\ &\quad + O(\Delta t^2). \end{aligned} \quad (\text{A } 3)$$

Divide both sides of equation (A 3) by Δt , and take the continuum limit as $\Delta t \rightarrow 0$, we have,

$$\frac{d \log Y_0(t)}{dt} = [a_i - \beta(t)] \frac{d \log N_i(t)}{dt} - \beta'(t) \log N_i(t). \quad (\text{A } 4)$$

The growth rates η and γ , as defined in the main text, follow $\eta = d \log Y_0(t)/dt$, and $\gamma_i = d \log N_i(t)/dt$. Plug these relationships into equation (A 4), we have,

$$\eta = [a_i - \beta(t)]\gamma_i - \beta'(t) \log N_i(t). \quad (\text{A } 5)$$

Rearrange terms, we have,

$$a_i = \beta(t) + \frac{\eta}{\gamma_i} - \beta'(t) \frac{\log N_i(t)}{\gamma_i}. \quad (\text{A } 6)$$

With time-independent β , the derivative $\beta'(t)=0$. Then equation (A 6) simplifies to

$$a_i = \beta + \frac{\eta}{\gamma_i}, \quad (\text{A } 7)$$

confirming the results in equations (2.19) and (2.20) of the main text.

Appendix B. Biological examples

In contrast to urban systems, the temporal scaling of many biological systems follows quite different behaviour. The life-cycle growth of many organisms has been understood in terms of the total available metabolism, which scales with body size, partitioned into growth and maintenance terms [59]. For mammals the total metabolic power scales with body size following an exponent of $\beta=3/4$, while the maintenance metabolism has an exponent of 1 which implies an asymptotic size. Taken together these two features imply that the growth rate of an organism is given by

$$\frac{dN_i}{dt} = \frac{1}{E_c} (B_0 N_i^\beta - B_c N_i), \quad (\text{B } 1)$$

where for organisms N_i is the number of cells in the body, E_c is the energy to create a cell, B_0 is the normalization constant of metabolic rate, and B_c is the metabolic rate to support a cell. In this framework organisms of any size, and at any point along their temporal growth trajectory, follow $B = B_0 m^\beta / m_c^\beta$ for metabolic rate (m_c is the mass of a cell), and the resulting growth trajectories are consistent with data.

It is interesting to consider the downstream effects of this model on other features. For example, if we consider the normalized biomass production rate as our $Y = (1/N_i)(dN_i/dt)$ of interest then it is important to note that the organismal growth equation does not reflect pure temporal scaling. For early times the temporal scaling will follow $N_i^{\beta-1}$, but eventually this term is overtaken by B_c because if $\beta < 1$ the organism asymptotes in size. Temporal scaling is only appropriate to consider over a certain range of time during early organism growth.

The asymptotic size ($\beta < 1$) is found at $dN_i/dt = 0$ and is given by $N_{\max} = (B_0/B_c)^{1/(1-\beta)}$, which allows us to rewrite

the normalized biomass production equation as

$$Y = \frac{B_0}{E_c} N_i^{\beta-1} \left[1 - \left(\frac{N_i}{N_{\max}} \right)^{1-\beta} \right]. \quad (\text{B } 2)$$

For both cross-sectional and temporal scaling, it is convenient to consider the system size at a particular fraction of the asymptotic size, $N_i = \varepsilon N_{\max}$, in which case we have

$$Y = \frac{B_0}{E_c} \varepsilon^{\beta-1} N_{\max}^{\beta-1} [1 - \varepsilon^{1-\beta}]. \quad (\text{B } 3)$$

The cross-sectional scaling can be found by considering a fixed value of ε in which case $Y \propto N_{\max}^{\beta-1}$ for any value of ε . Thus the temporal and cross-sectional scaling relationships for Y have identical exponents for animals, but only in the case of early organism growth after which temporal scaling breaks down.

Appendix C. Figure 3 with all observations

For clarity of presentation, we omitted a number of outliers in figure 3. Here, we show all observations in the analysis in figure 4.

Figure 4. Estimates of the temporal exponent via equation (2.13) compared to direct ordinary least squares minimization. Showing all observations, including outliers that were not included in figure 3. (Online version in colour.)

References

1. Barenblatt GI. 2003 *Scaling*, vol. 34. Cambridge, UK: Cambridge University Press.
2. Landau LD, Lifshitz E. 1980 *Course of theoretical physics*. Oxford, UK: Pergamon Press Ltd.
3. Cody ML et al. 1975 *Ecology and evolution of communities*. Cambridge, MA: Harvard University Press.
4. Hamilton MJ, Milne BT, Walker RS, Burger O, Brown JH. 2007 The complex structure of hunter-gatherer social networks. *Proc. R. Soc. B* **274**, 2195–2203. (doi:10.1098/rspb.2007.0564)
5. Bettencourt LMA, Lobo J, Helbing D, Kühnert C, West GB. 2007 Growth, innovation, scaling, and the pace of life in cities. *Proc. Natl Acad. Sci. USA* **104**, 7301–7306. (doi:10.1073/pnas.0610172104)
6. Batty M. 2007 *Cities and complexity: understanding cities with cellular automata, agent-based models, and fractals*. Cambridge, MA: The MIT Press.
7. Kardar M. 2007 *Statistical physics of particles*. Cambridge, UK: Cambridge University Press.
8. West GB, Brown JH, Enquist BJ. 1997 A general model for the origin of allometric scaling laws in biology. *Science* **276**, 122–126. (doi:10.1126/science.276.5309.122)
9. West GB, Brown JH, Enquist BJ. 1999 The fourth dimension of life: fractal geometry and allometric scaling of organisms. *Science* **284**, 1677–1679. (doi:10.1126/science.284.5420.1677)
10. Harte J. 2011 *Maximum entropy and ecology: a theory of abundance, distribution, and energetics*. Oxford, UK: OUP.
11. MacArthur RH. 1965 Patterns of species diversity. *Biol. Rev.* **40**, 510–533. (doi:10.1111/j.1469-185X.1965.tb00815.x)

12. MacArthur RH, Wilson EO. 2001 *The theory of island biogeography*, vol. 1. Princeton, NJ: Princeton University Press.
13. Rosindell J, Hubbell SP, Etienne RS. 2011 The unified neutral theory of biodiversity and biogeography at age ten. *Trends Ecol. Evol.* **26**, 340–348. (doi:10.1016/j.tree.2011.03.024)
14. Simon HA. 1955 On a class of skew distribution functions. *Biometrika* **42**, 425–440. (doi:10.1093/biomet/42.3-4.425)
15. Barabási A-L, Albert R. 1999 Emergence of scaling in random networks. *Science* **286**, 509–512. (doi:10.1126/science.286.5439.509)
16. Bettencourt LMA. 2013 The origins of scaling in cities. *Science* **340**, 1438–1441. (doi:10.1126/science.1235823)
17. Bettencourt LMA. 2014 Impact of changing technology on the evolution of complex informational networks. *Proc. IEEE* **102**, 1878–1891. (doi:10.1109/JPROC.2014.2367132)
18. Daepf MI, Hamilton MJ, West GB, Bettencourt LMA. 2015 The mortality of companies. *J. R. Soc. Interface* **12**, 20150120. (doi:10.1098/rsif.2015.0120)
19. Arbesman S, Kleinberg JM, Strogatz SH. 2009 Superlinear scaling for innovation in cities. *Phys. Rev. E* **79**, 016115. (doi:10.1103/PhysRevE.79.016115)
20. Pan W, Ghoshal G, Krumme C, Cebrian M, Pentland A. 2013 Urban characteristics attributable to density-driven tie formation. *Nat. Commun.* **4**, 1961. (doi:10.1038/ncomms2961)
21. Yang VC, Papachristos AV, Abrams DM. 2019 Modeling the origin of urban-output scaling laws. *Phys. Rev. E* **100**, 032306. (doi:10.1103/PhysRevE.100.032306)
22. Gomez-Lievano A, Patterson-Lomba O, Hausmann R. 2017 Explaining the prevalence, scaling and variance of urban phenomena. *Nat. Hum. Behav.* **1**, 0012. (doi:10.1038/s41562-016-0012)
23. Bettencourt LMA, Lobo J, Strumsky D, West GB. 2010 Urban scaling and its deviations: revealing the structure of wealth, innovation and crime across cities. *PLoS ONE* **5**, e13541. (doi:10.1371/journal.pone.0013541)
24. Alves LG, Ribeiro HV, Lenzi EK, Mendes RS. 2013 Distance to the scaling law: a useful approach for unveiling relationships between crime and urban metrics. *PLoS ONE* **8**, e69580. (doi:10.1371/journal.pone.0069580)
25. Sahasranaman A, Bettencourt LMA. 2019 Urban geography and scaling of contemporary Indian cities. *J. R. Soc. Interface* **16**, 20180758. (doi:10.1098/rsif.2018.0758)
26. Bettencourt LMA, Lobo J. 2016 Urban scaling in Europe. *J. R. Soc. Interface* **13**, 20160005. (doi:10.1098/rsif.2016.0005)
27. Sveikauskas L. 1975 The productivity of cities. *Q. J. Econ.* **89**, 393–413. (doi:10.2307/1885259)
28. O'sullivan A. 2011 *Urban economics*. New York, NY: McGraw-Hill/Irwin.
29. Ribeiro FL, Meirelles J, Ferreira FF, Neto CR. 2017 A model of urban scaling laws based on distance dependent interactions. *R. Soc. open sci.* **4**, 160926. (doi:10.1098/rsos.160926)
30. Molinero C, Thurner S. 2019 How the geometry of cities explains urban scaling laws and determines their exponents. (<http://arxiv.org/abs/1908.07470>).
31. Gomez-Lievano A, Youn H, Bettencourt LMA. 2012 The statistics of urban scaling and their connection to Zipf's law. *PLoS ONE* **7**, e40393. (doi:10.1371/journal.pone.0040393)
32. Depersin J, Barthelemy M. 2018 From global scaling to the dynamics of individual cities. *Proc. Natl Acad. Sci. USA* **115**, 2317–2322. (doi:10.1073/pnas.1718690115)
33. Bettencourt LMA, Lobo J, Strumsky D. 2007 Invention in the city: increasing returns to patenting as a scaling function of metropolitan size. *Res. Policy* **36**, 107–120. (doi:10.1016/j.respol.2006.09.026)
34. Ciccone A. 2002 Agglomeration effects in Europe. *Eur. Econ. Rev.* **46**, 213–227. (doi:10.1016/S0014-2921(00)00099-4)
35. Duranton G, Puga D. 2004 Micro-foundations of urban agglomeration economies. In *Handbook of regional and urban economics*, vol. 4, pp. 2063–2117. Elsevier.
36. Jones CI. 1999 Growth: with or without scale effects? *Am. Econ. Rev.* **89**, 139–144. (doi:10.1257/aer.89.2.139)
37. Jones CI. 2005 Growth and ideas. In *Handbook of economic growth*, vol. 1 (eds P Aghion, SN Durlauf), pp. 1063–1111. Elsevier.
38. Romer PM. 1990 Endogenous technological change. *J. Political Econ.* **98**, S71–S102. (doi:10.1086/261725)
39. Barthelemy M. 2016 *The structure and dynamics of cities: urban data analysis and theoretical modeling*. Cambridge, UK: Cambridge University Press.
40. Ortman SG, Cabaniss AH, Sturm JO, Bettencourt LMA. 2014 The pre-history of urban scaling. *PLoS ONE* **9**, e87902. (doi:10.1371/journal.pone.0087902)
41. Ortman SG, Cabaniss AH, Sturm JO, Bettencourt LMA. 2015 Settlement scaling and increasing returns in an ancient society. *Sci. Adv.* **1**, e1400066. (doi:10.1126/sciadv.1400066)
42. Ortman SG, Coffey GD. 2017 Settlement scaling in middle-range societies. *Am. Antiq.* **82**, 662–682. (doi:10.1017/aaq.2017.42)
43. Cesaretti R, Lobo J, Bettencourt LMA, Ortman SG, Smith ME. 2016 Population–area relationship for medieval European cities. *PLoS ONE* **11**, e0162678. (doi:10.1371/journal.pone.0162678)
44. Hanson JW. 2016 *An urban geography of the Roman world, 100 BC to AD 300*, vol. 18. Oxford, UK: Archaeopress oxford.
45. Brelsford C, Lobo J, Hand J, Bettencourt LMA. 2017 Heterogeneity and scale of sustainable development in cities. *Proc. Natl Acad. Sci. USA* **114**, 8963–8968. (doi:10.1073/pnas.1606033114)
46. Bettencourt LMA. Submitted. Urban growth and the emergence statistics of cities.
47. Keuschnigg M. 2019 Scaling trajectories of cities. *Proc. Natl Acad. Sci. USA* **116**, 201906258.
48. Batty M, Longley PA. 1994 *Fractal cities: a geometry of form and function*. New York, NY: Academic Press.
49. Pumain D, Paulus F, Vacchiani-Marcuzzo C, Lobo J. 2006 An evolutionary theory for interpreting urban scaling laws. *Cybergeo: Eur. J. Geogr.* **2006**, 1–20.
50. Lämmer S, Gehlsen B, Helbing D. 2006 Scaling laws in the spatial structure of urban road networks. *Physica A* **363**, 89–95. (doi:10.1016/j.physa.2006.01.051)
51. Rybski D, Reusser DE, Winz AL, Fichtner C, Sterzel T, Kropp JP. 2017 Cities as nuclei of sustainability? *Environ. Plan. B* **44**, 425–440. (doi:10.1177/0265813516638340)
52. Wooldridge JM. 2002 *Econometric analysis of cross section and panel data*. Cambridge, MA: The MIT Press.
53. Baltagi BH. 2013 *Econometric analysis of panel data*. Hoboken, NJ: Wiley.
54. Hsiao C. 2014 *Analysis of panel data*. Cambridge, UK: Cambridge University Press.
55. Anderson P. 1972 More is different. *Science* **177**, 393–396. (doi:10.1126/science.177.4047.393)
56. Alonso W *et al.* 1964 *Location and land use. Toward a general theory of land rent*. Cambridge, MA: Harvard University Press.
57. Bristow DN, Kennedy CA. 2013 The energy for growing and maintaining cities. *Ambio* **42**, 41–51. (doi:10.1007/s13280-012-0350-x)
58. Lobo J, Bettencourt LMA, Strumsky D, West GB. 2013 Urban scaling and the production function for cities. *PLoS ONE* **8**, e58407. (doi:10.1371/journal.pone.0058407)
59. West GB, Brown JH, Enquist BJ. 2001 A general model for ontogenetic growth. *Nature* **413**, 628. (doi:10.1038/35098076)
60. Kempes CP, Dutkiewicz S, Follows MJ. 2012 Growth, metabolic partitioning, and the size of microorganisms. *Proc. Natl Acad. Sci. USA* **109**, 495–500. (doi:10.1073/pnas.1115585109)