https://doi.org/10.1140/epjds/s13688-019-0203-y

EPJ Data Science
a SpringerOpen Journal

REGULAR ARTICLE Open Access

Testing Heaps' law for cities using administrative and gridded population data sets

Filippo Simini[1*] and Charlotte James[1]

*Correspondence:
f.simini@bristol.ac.uk
[1] Department of Engineering Mathematics, University of Bristol, Bristol, UK

Abstract

Since 2008 the number of individuals living in urban areas has surpassed that of rural areas and in the next decades urbanisation is expected to further increase, especially in developing countries. A country's urbanisation depends both on the distribution of city sizes, describing the fraction of cities with a given population (or area), and the overall number of cities in the country. Here we present empirical evidence suggesting the validity of Heaps' law for cities: the expected number of cities in a country is only a function of the country's total population (or built-up area) and the distribution of city sizes. This implies the absence of correlations in the spatial distribution of cities. We show that this result holds at the country scale using the official administrative definition of cities provided by the Geonames dataset, as well as at the local scale, for areas of 128 × 128 km^2 in the United States, using a morphological definition of urban clusters obtained from the Global Rural-Urban Mapping Project (GRUMP) dataset. We also derive a general theoretical result applicable to all systems characterised by a Zipf distribution of group sizes, which describes the relationship between the expected number of groups (cities) and the total number of elements in all groups (population), providing further insights on the relationship between Zipf's law and Heaps' law for finite-size systems.

Keywords: Heaps' law; Zipf's law; Cities; Urbanisation; Scaling

1 Introduction

The increase of urbanisation rates, generally defined as the increase of the proportion of people living in urban areas or the proportion of buildings belonging to urban agglomerations [1], is a trend that has happened in waves throughout human history, with a dramatic acceleration in the last 300 years [2]. In 2015, 56% of China's population lived in cities, a figure that has more than doubled compared to 26% of 1990. The Ministry of Housing and Urban-Rural Development estimates that by 2025 300M Chinese now living in rural areas will move into cities. State spending is planned on new houses, roads, hospitals, schools, which could cost up to 600 billion US dollars a year. A great rate of urbanisation is also expected in Sub-Saharan African countries. As a result, by 2030 it is estimated that the world's population will have increased by over 1 billion people most of whom will dwell in the rapidly growing cities of Asia and Africa [3]. Recent studies show that, on average, urban land is expanding at twice the urban population growth rate, resulting in a decrease of urban population density with time [4].

A quantitative understanding of the mechanisms that drive urbanisation is important for helping governments and decision makers to plan investments in order to achieve sustainable urban planning and growth. These decisions will have a huge impact on the lives of millions of people, the economy and the environment. Urbanisation can happen in two ways: diffusion (or sprawl) and aggregation. Diffusion corresponds to existing cities growing and increasing in size because of either net migration from rural areas or a greater rate of natural increase (i.e. birth rate minus death rate) in urban areas. Aggregation corresponds to new villages and towns being created in rural areas that were previously considered non-urbanised. In order to properly characterise urbanisation patterns we should consider both aspects: the distribution of city sizes, describing the size and growth of existing cities, and the overall number of cities, describing the abundance and formation of new urban areas.

The distribution of city sizes is a broad and heterogeneous distribution. Ranking cities by population, it has been observed [5–7] that the population of the i-th largest city of a country is approximately equal to the population of the largest city divided by i, i.e. a city's rank is inversely proportional to its population. In other words, the fraction of cities with population larger than x follows Zipf's law, $P(> x) \sim x^{-\alpha}$, with $\alpha \simeq 1$. Previous studies have shown how Zipf's law can originate from various models based on cluster growth and aggregation [8–11], the interplay between multiplication and diffusion processes [12], preferential migration to large aggregates [13], pairwise interactions between individuals [14] and proportionate random growth [15–17], or Gibrat's law [18, 19].

Compared to the great efforts made to characterise the distribution of city sizes both empirically and theoretically, much less work has been done to answer the other fundamental question about the urbanisation process: What determines the number of cities in a country? In this paper we empirically investigate the relationships between the number of cities in a region and some of the region's properties, such as the region's total population and built-up area. In particular, we consider how the total population (or the total built-up area) of a region affects the number of cities. This is analogous to Heaps' Law in linguistics [20, 21], which describes the empirical scaling relationship between the number of distinct words, W, in a document and the total number of words in the document (or text length), N: $W \sim N^{\gamma}$, where $\gamma \leq 1$ is the Heaps exponent.

Previous research has shown that Zipf's law and Heaps' law often appear together, suggesting that the presence of Zipf's law implies Heaps' law. Considering the probability density function (PDF) corresponding to Zipf's Law, $P(x) \sim x^{-1-\alpha}$, it can be shown [22] that Heaps' exponent γ is related to Zipf's exponent α: $\gamma = \alpha$ if $\alpha < 1$, and $\gamma = 1$ otherwise. However, this relationship does not necessarily hold for spatially extended systems, such as cities, because evidence of Zipf's law at the country (global) scale does not necessarily imply the presence of Zipf's law and Heaps' law at the regional (local) scale. In fact, even if Zipf's law for the distribution of city sizes holds globally at the level of countries, it might not hold locally at smaller spatial scales if correlations in the spatial distribution of urban clusters are present. This would be the case, for example, if urban clusters were spatially aggregated by size, so that it is more common to find clusters of similar sizes close to each other compared to the case in which clusters are randomly distributed among the regions, irrespective of their size. The overall (global) distribution of cluster sizes would not change and still be a power-law, but the size distributions in the regions would not follow Zipf's law anymore and as a consequence Heaps' law would not hold. Indeed, this is what happens in ecological systems, where macro-ecological statistical patterns of species distribution and abundance display a strong dependence on the spatial scale considered [23]. One of the most relevant statistics used to characterise the degree of biodiversity of ecosystems is the species-area relationship (SAR), which measures the number of different species expected to be found in areas of increasing size. Since the density of individuals per unit area is constant, the SAR is the equivalent of Heaps' law for ecosystems, as it measures the relationship between a region's total population and the expected number of different groups of individuals in the region, where here groups correspond to species instead of cities. Empirical measurements of the SAR show a different functional behaviour as the region's area increases, and this is due to the fact that the shape of the distribution of species sizes, called the relative species abundance, depends on the spatial scale considered. While there are various studies on Heaps' law in linguistics and SAR in ecology, to the best of our knowledge there is no thorough empirical analysis of Heaps' law in urban systems. The aim of this paper is to precisely fill this gap and to investigate the validity of Heaps' laws for cities.

There is another reason to investigate the relationship between Zipf's and Heaps' laws for cities. Zipf's law for the distribution of city sizes usually holds only for the tail of the distribution, however the fact that in a region the distribution of city sizes has a power-law tail does not give any information regarding the relationship between the number of cities in the region and its total population. In other words, when Zipf's law holds only for large cities, there is no guarantee that Heaps' law holds as well. To understand this, consider a region in which city sizes follow Zipf's law. If the population of each city is doubled and hence the total population of the region is also doubled, yet no new cities are created, Zipf's law will still be present, albeit with a larger scale parameter (i.e. the minimum city size is doubled). However, Heaps' law will not hold in this case, because the total population, N, is doubled, but the number of cities, C, has not changed.

In this paper, we use a dataset on the population and location of cities globally to assess if Heaps' law holds for all countries in all continents (except Australia and Antartica), and to test the predicted relationship between Heaps' and Zipf's exponents. Cities can be defined in many different ways and various relevant properties of urban agglomerations, including the scaling relationships between population size and urban indicators such as area of roads and number of patents, depend on the method used to define cities [24, 25]. In particular, the relationship between the number of cities in a region and the region's total population, i.e. Heaps' law, can also depend on the definition of city considered. To understand how Heaps' law depend on the definition of city, we use a second dataset of the spatial distribution of population in the United States that allows us to consider various definitions of urban clusters and provide additional support to our results.

2 Analytical results

Our aim is to measure the relationship between the distribution of city sizes and the expected number of cities in various regions worldwide in order to understand if these patterns can be described by a simple null model which assumes that cities are independently and randomly distributed in space. The only assumption of this null model is the global distribution of city sizes, i.e. Zipf's law, which is used to populate an initially empty region.

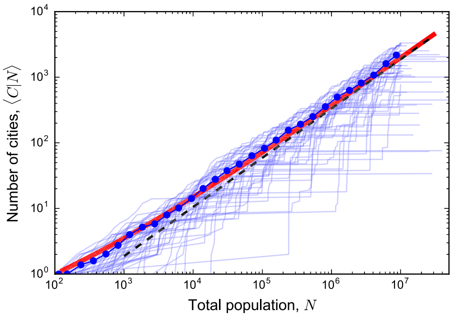

Figure 1 Theoretical prediction. Relationship between the expected number of cities and the total population for the null model, which assumes a Zipf law with exponent $\alpha = 0.75$ and minimum city size $m = 10^2$ for the distribution of city sizes and no spatial correlations between cities. The thin blue lines are 100 realisations of the model and the blue circles denote the average number of cities, $\langle C|N\rangle$, for a fixed value of the total population, N. The solid red line is the theoretical prediction of Equation (1) and the dashed black line is Heaps' law, $\langle C|N\rangle \sim N^{\alpha}$, which holds in the limit of very large population $N \gg m$

One realisation of the model consists in drawing cities from this global distribution until a given target total size of the region, N, is reached. As soon as the sum of the city sizes becomes larger than N the drawing stops and the number of cities, C, corresponds to the number of drawings in this realisation. Repeating this process and averaging over many realisations it is possible to estimate the expected number of cities for a fixed target total population N, $\langle C|N\rangle$. Varying N, one can study the dependence of the expected number of cities on the region's total population and assess the validity of Heaps' law.

A simple calculation shows that, under the assumptions of this null model, Heaps' law holds asymptotically for very large populations, as reported in the literature [22]. Additionally, our calculation allows us to derive a more accurate formula for the relationship between the number of cities and the total population, which is valid for smaller populations as well. Let us consider the probability to find a city with population x within a group of C cities with total population N, $p(x|C,N)$. We can use this probability to compute the average population of such a group of cities, which is equal to N/C: $N = C \cdot \langle x|C,N\rangle$, where $\langle x|C,N\rangle$ denotes the conditional expectation of x given C and N. Multiplying by the probability to find C cities in a region with total population N, $p(C|N)$, on both sides and integrating with respect to C we get $N = \int \mathrm{d}Cp(C|N)C \int \mathrm{d}xxp(x|C,N)$. If the probability $p(x|C,N)$ can be considered independent of the number of cities when $C \gg 1$, i.e. $p(x|C,N) \approx p(x|N)$ and thus $\langle x|C,N\rangle \approx \langle x|N\rangle$, then the expected number of cities in a region with population N is $\langle C|N\rangle \approx N/\langle x|N\rangle$. Using the assumption that city sizes follow Zipf's law with exponent $\alpha < 1$ and given that the maximum city size cannot be larger than the region's total population, we can write $p(x|N) = \alpha/(m^{-\alpha} - N^{-\alpha})x^{-1-\alpha}\Theta(N - x)$, where m is the minimum city size and Θ is the Heaviside step function. From this, we obtain the following equation relating the average number of cities and the total population:

$$\langle C|N\rangle \approx \frac{1-\alpha}{\alpha} \cdot N \frac{(m^{-\alpha} - N^{-\alpha})}{(N^{1-\alpha} - m^{1-\alpha})}, \qquad (1)$$

where m represents the minimum city size. Note that $\langle C|N\rangle$ is a function of the ratio N/m. When the region's population is very large, $N \gg m$, Equation (1) can be approximated as $\langle C|N\rangle \sim (N/m)^{\alpha}$, i.e. we obtain Heaps' law. Figure 1 shows 100 realisations of the null model and the theoretical prediction given by Equation (1) for $\alpha = 0.75$.

3 Heaps' law for countries

Our first goal is to assess if Heaps' law holds for all countries in the four most populated continents. To this end, we analyse the relationship between the number of cities in each

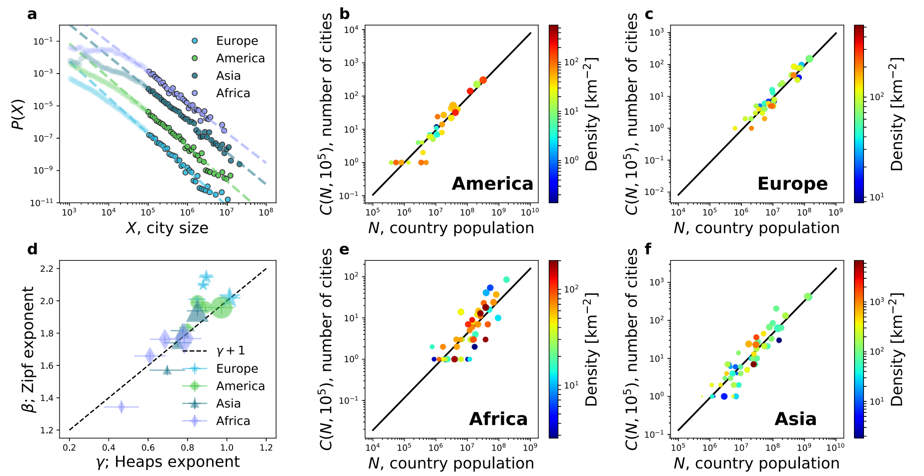

Figure 2 Heaps' law for countries. (**a**), Zipf's Law: PDF (y-axis) of city sizes X (x-axis) for all cities in Europe, America, Asia and Africa. The darker regions correspond to cities with population $X > 10^5$, above which the distributions are a power law with exponent $\beta = 1 + \alpha$ given in Table 1. The dashed lines correspond to $y = x^{-\beta}$. Distributions have been shifted in the y-axis for clarity.(**b**)–(**c**) and (**e**)–(**f**), Heaps' law for America, Europe, Africa and Asia. The following information is displayed for each country: population (x-axis), number of cities with more than 100 k inhabitants (y-axis), logarithm of the area (marker size) and population density (color). The black line is a power law fit of the scaling relationship between the number of cities and the total population; Heaps exponents γ are reported in Table 1. (**d**), The exponent of the Zipf PDF, β (y-axis) and the corresponding exponent γ of Heaps' law for Europe, America, Asia and Africa. Marker size corresponds to the minimum city population used in determining the values of γ and β: values used are $10^3, 5 \times 10^3, 10^4$ and 10^5, where 10^3 is represented by the smallest marker and 10^5 by the largest. Increasing the minimum city population corresponds to a decrease in the amount of data used to fit γ and β: for a given continent, each point represents the same data set but restricted to a different range of X values. The black dashed line corresponds to the relationship between the exponents, $\beta = \gamma + 1$

country and the country's total population. Since most countries have large populations, we expect that data should follow the asymptotic form of Heaps' law, $C \sim N^{\alpha}$, if the assumptions of our null model hold. To test this prediction, in Fig. 2(a) we fit a power law to the tail of the empirical distribution of city sizes for each continent, obtaining the Zipf PDF exponent $\beta = \alpha + 1$. Then we fit a power law to the scatter plot between the number of cities in each country and the country's total population, Figs. 2(b)–(c), 2(e)–(f), obtaining the Heaps exponent γ. Finally, we check if the value of α is equal to γ for each of the continents, Fig. 2(d).

The dataset used to perform this analysis is the *Geonames* dataset [26], which consists of the population and geographic location of all cities with more than 1000 inhabitants worldwide. Data on the area and population of all countries was obtained from Worldbank [27].

Zipf's law We consider four continents: Africa, Asia, Europe and America (North and South). We find that the distribution of city sizes follows Zipf's Law for cities above a minimum population of $\sim 10^5$ as shown in Fig. 2(a), where the darker points correspond to points where population is larger than 10^5. The exponents of the PDFs, $\beta = 1 + \alpha$, for all continents are displayed in Table 1, along with their errors. We observe that while America and Europe both satisfy Zipf's law with exponents compatible with $\alpha = 1$ within errors, Asia and Africa have exponents significantly smaller than 1.

Table 1 Exponents of Zipf's Law and Heaps' Law. Column 1 displays $\beta = 1 + \alpha$; the PDF exponent of Zipf's Law with corresponding standard deviation displayed in column 2 for each of the four continents. Values were calculated by fitting a line to the PDF of city sizes X for each continent, starting from a minimum population of $X = 10^5$ (see Fig. 2(a)). Column 3 displays γ; the exponent of Heaps' Law with corresponding standard deviation displayed in column 4 for each of the four continents. Values of γ and the error were obtained by fitting a line to the logarithm of Heaps' Law; $\log(C) = \gamma \log(N)$, for each continent (see Figs. 2(b)–(c) and (e)–(f)). Exponents β and γ were fit to the data using non-linear least squares

Continent	β	σ_β	γ	σ_γ
Africa	1.77	0.05	0.78	0.09
Asia	1.94	0.04	0.85	0.05
America	1.96	0.04	0.97	0.07
Europe	2.02	0.05	1.01	0.06

Heap's law We analyse the relationship between the number of cities in a country, C, and the country's population, N, for all African, Asian, American and European countries (Fig. 2(b)–(c) and (e)–(f)). We fit a power law to these data and obtain the Heaps exponents γ reported in Table 1. To test the validity of Heaps' Law for cities, we assess the extent to which Heaps' exponents γ are equal to the exponents $\alpha = \beta - 1$ from Zipf's Law. In Fig. 2(d) we plot γ (x-axis) vs β (y-axis) for different values of the minimum city population, where the exponents are fit to cities with population greater than this minimum population. The best fit values of Heaps exponent γ and Zipf PDF exponent $\beta = \alpha + 1$ are compatible with the relationship $\gamma = \alpha$ for all the continents (see Table 1), supporting the validity of the null model. The theoretical relationship $\gamma = \beta - 1$ (black dashed line) is better satisfied when we consider large cities ($> 10^5$), whereas there are significant deviations for small values of minimum population. This is explained by the fact that the distributions of city sizes are not pure power laws, but there are deviations from Zipf's law for small cities (see Fig. 2(a)).

Heaps' law in the United States Heaps' law is further confirmed by considering more homogeneous sets of regions, like the United States in Fig. 3(a). There is clear evidence that the number of cities grows proportionally with the state population, whilst there is a small or indirect relationship between the number of cities and the state's area or population density: in the United States, the cities-population, cities-area, cities-density correlation coefficients are 0.95, 0.04, and –0.08 respectively. In Fig. 3(b) we plot the number of cities with more than X inhabitants in each United States state, $C(N,X)$, as a function of the ratio N/X for values of X ranging from 5000 to 5,000,000 inhabitants. All points collapse on a straight line, confirming that the equation $C(N,X) \sim N/X$ holds for several orders of magnitude of N and X. This is confirmed for the four continents as well (see see Additional file 1). We also find evidence of the validity of Heaps' law throughout time in the state of Iowa, United States. Historical data shows that between 1850 and 2000 the number of incorporated places (i.e. self-governing cities, towns, or villages) grew at the same rate as the state population (Fig. 3(c)).

Spatial distribution of cities We use the *Geonames* database to test another assumption of the null model about the absence of spatial correlations in the distribution of city sizes. If cities are randomly and uniformly distributed in space, it follows that the average distance to the closest city for cities with more than X inhabitants, $\langle d_c \rangle$, is proportional to the square

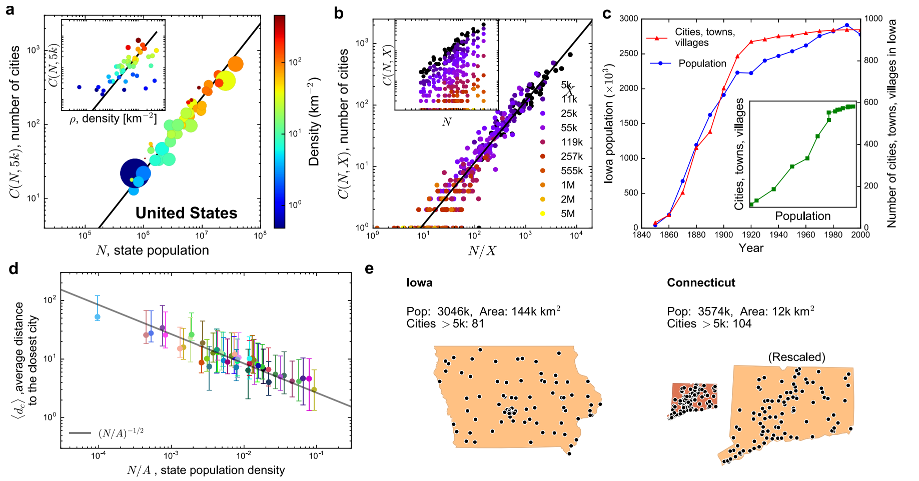

Figure 3 Consequences of Heaps' law. (**a**), The number of cities with more than 5000 inhabitants in the Unites States is proportional to the state's population, corr(C, N) = 0.95. The correlations with area (0.04) and population density (–0.08, see inset) are negligible, as illustrated by the following pairs of states with similar area or density and very different number of cities: Alaska (A = 1.5M km^2, $C(5k)$ = 22) vs Texas (A = 0.7M km^2, $C(5k)$ = 392), and Rhode Island (ρ = 393 km^{-2}, $C(5k)$ = 35) vs New Jersey (ρ = 467 km^{-2}, $C(5k)$ = 316). (**b**), Combining the result from panel (**a**) with Zipf's law it is possible to estimate the number of cities with more than X inhabitants in a country with population N as $C(N,X) \sim N/X$; Heaps' law. As a consequence, the scattered cloud of points resulting when plotting $C(N,X)$ against N for various X's in the range $5 \cdot 10^3 - 5 \cdot 10^6$ (inset) collapses on a straight line when $C(N,X)$ is plotted against the ratio N/X. (**c**), Historical records of the number of incorporated places (C, red triangles) and the state population (N, blue circles) in Iowa from 1850 to 2000 (source: State library of Iowa, state data center). The similar growth rates of C and N entail the validity of Heaps' law $C \sim N$ during the 150-year period (inset). (**d**), The average distance to the closest city in the United States scales as the inverse of the square root of the state's population density (here all cities with more than 5000 inhabitants are considered). The asymmetric error bars denote the standard deviations above and below the average. (**e**), Illustration of the relationships between total population, number of cities, and their average distance in Iowa and Connecticut. In agreement with Heaps' law, $C \sim N^{\alpha}$, Iowa and Connecticut have similar populations and a similar number of cities with more than 5000 inhabitants, despite Connecticut having one-twelfth the area of Iowa. In agreement with Equation (2), cities in Connecticut are closer than cities in Iowa because of the higher population density in Connecticut. By rescaling distances such that Connecticut's area becomes equal to Iowa's area, the two states would have the same population density and consequently the same average distance between cities

root of X and inversely proportional to the square root of the region's population density, $\rho \equiv N/A$:

$$\langle d_c \rangle \sim \sqrt{X/\rho}. \tag{2}$$

In fact, when cities are randomly and uniformly distributed in space, the average number of cities in a region with uniform population density (if measured on a length scale larger than the average distance between cities) is proportional to the region's area, or equivalently the density of cities scales as $\chi \sim C/A$. Combining this result with Heaps' law, $C \sim N/X$, and observing that the average distance to the closest city, $\langle d_c \rangle$, scales as the inverse of the square root of the density of cities, we obtain the result of Equation (2): $\langle d_c \rangle \sim 1/\sqrt{\chi} \sim \sqrt{A/C} \sim \sqrt{X(A/N)} \sim \sqrt{X/\rho}$.

Figure 3(d) shows the average distance to the closest city with more than X = 5000 inhabitants for the United States' states as a function of the state population density, and confirms the scaling behaviour predicted by Equation (2). An identical analysis using African, Asian, American and European countries provides further support for this scaling behavior (see see Additional file 1).

This finding supports some of the conclusions of the Central Place Theory of human geography [28, 29], whilst disproving others. On the one hand, it is true that for regions with a given population density the larger the cities are, the fewer in number they will be, and the greater the distance, i.e. increasing X in Equation (2) results in a greater average distance $\langle d_c \rangle$. On the other hand, the average distance between cities of a given size X is not the same for all the states, but depends on the state's population density: cities of a given size are closer in densely populated states than in sparsely populated ones, i.e. for a fixed city size X and state area A the distance between cities decreases as the inverse square root of the state population, N (see Fig. 3(d)).

4 Heaps' law at short spatial scales

To understand how Heaps' law depends on the definition of city, we analyse data from the Global Rural-Urban Mapping Project [30] (GRUMPv1) consisting of estimates of the residential population of the United States for the year 2000 at a resolution of 30 arcseconds ($\sim$ 1 km).

Extraction of urban clusters In the GRUMP data the spatial distribution of population is represented as a matrix, whose elements denote the estimated number of individuals resident within each of the grid cells. We apply a city clustering algorithm [10] (CCA) to the GRUMP data and define cities as spatial clusters of neighbouring grid cells with population over a given threshold, m, which also corresponds to the minimum cluster population. We vary the parameter m over the interval [10–600] persons per km^2, clustering adjacent cells with population above the threshold m. As a reference, the official definition of urban area adopted by the United States census considers values of m between 193 and 386 people per square kilometer [31]. In the range of m considered, the numbers and sizes of clusters obtained with the CCA are very different. Panels b-d of Fig. 4 show the clusters within a square region in the Midwestern United States (Fig. 4(a)) for $m = 28, 129$, and 599. Both the number and areas of clusters decrease as m increases and some large clusters split into multiple smaller clusters.

Global distribution of areas and populations of urban clusters Additionally, the gridded population data allows us to consider the area of urban clusters [32–34], a, as the relevant size variable, alternatively to population, x. Indeed, the distribution of urban areas is also known to follow Zipf's law with exponent $\alpha \simeq 1$, hence our null model predicts that the number of clusters is given by Equation (1), where N now denotes the total urbanised area and α the exponent of the distribution of city areas. The area and population of urban clusters are strongly correlated variables. The expansion of urban areas can be characterised by measuring the scaling relationship between the area, a, and population, x, of the clusters. We use the gridded population data to measure urban sprawl for different definitions of city, i.e. different values of the CCA parameter m (see Fig. 4(e)). We observe that the scaling relationship between a and x has the following dependence on the minimum population parameter m:

$$a \sim (x/m)^{\omega(m)}. \tag{3}$$

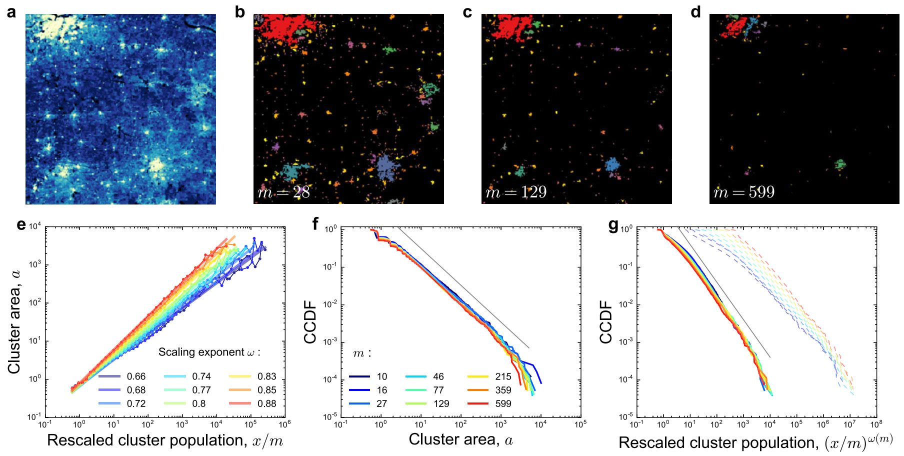

Figure 4 Scaling properties of urban clusters. (**a**), Portion of the GRUMP dataset representing the distribution of population in a region of the Midwestern United States. The color denotes the logarithm of the population: light yellow for high population, dark blue for low population. (**b**)–(**d**), Urban clusters in the region depicted in panel (**a**) obtained applying the City Clustering Algorithm for different values of the minimum population parameter: $m = 28$ in (**b**), 129 in (**c**), 599 in (**d**). (**e**), Scaling relationship between area and population of clusters. The different colors denote different values of the CCA parameter m (see the legend of panel (**f**) for the m values). The points indicate the average area of clusters with a given rescaled population. Data are fit to the power law in Equation (3) and the legend reports the values of the scaling exponent $\boldsymbol{\omega}$ for the various m. (**f**), Counter Cumulative Distribution Functions of cluster areas. The different colors denote different values of the CCA parameter m (see the legend for the m values). The grey line is a power law with exponent –1 as a guide for the eye. (**g**), Counter Cumulative Distribution Functions of cluster populations, x (dashed curves), and rescaled populations, $(x/m)^{\omega(m)}$ (solid curves). The grey line is a power law with exponent –1 as a guide for the eye

Note that the area of a cluster scales with the ratio x/m, which represents the maximum area that a cluster of population x can have, given m. The scaling exponent ω depends on m. In particular, $\omega(m)$ is an increasing function of m, which grows from 0.66 to 0.88. The sublinear scaling ($\omega(m) < 1$ for all m) between a cluster's area and population implies an increase in the population density of large clusters: the population density scales as $x/a = x^{1-\omega}$, which is a growing function of x when $\omega < 1$. This result may support the hypothesis on the economies of scale in the use of urban space. In fact, in large clusters space is organised more efficiently than in small clusters, so that each square kilometre of land can host a larger number of individuals, hence increasing the cluster's population density [24]. Urban sprawl happens when the exponent ω has a large value, indicating a reduced efficiency in the utilisation of space as the size of clusters grows. The fact that ω increases with m means that the estimated urban sprawl is bigger when clusters are defined using a large m and smaller when m is small. The scaling relationship between area and population of clusters, Equation (3), implies that the Zipf exponents of the distributions of cluster areas and populations, α_a and α_x respectively, are not independent, but related by the equation $\alpha_x = \alpha_a \cdot \omega(m)$.

The empirical distributions of cluster areas for different values of the CCA parameter m, shown in Fig. 4(f), indicate that the Zipf exponent for the areas is $\alpha_a \simeq 1$, independent of m. The distributions of cluster populations, instead, have exponents that depend on m. If the populations are rescaled by m and elevated to the power of $\omega(m)$, the curves for different m collapse on the same power law with exponent $\alpha_a \simeq 1$, verifying the relationship $\alpha_x = \alpha_a \cdot \omega(m)$ (see Fig. 4(g)).

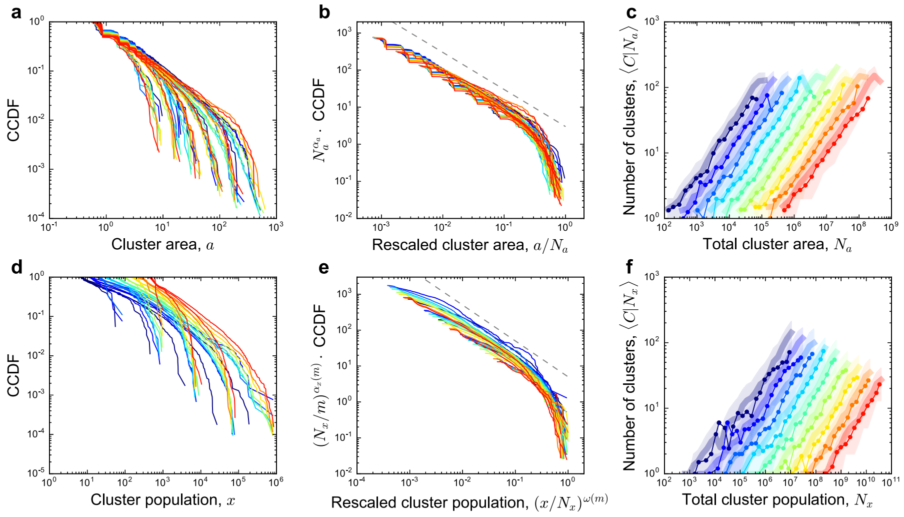

Figure 5 Zipf's law and Heaps' law for urban clusters. (**a**), Counter Cumulative Distribution Functions of areas of the clusters in all United States regions of 128 × 128 km^2 having urbanised area up to 5%. Regions are grouped in six groups according to their total urbanised area, N_a, and the CCDFs of each group are computed separately for the different values of the CCA parameter m (see the legend in Fig. 4(f) for the m values). (**b**), The CCDFs of panel (**a**) collapse on the same curve when the axes are properly rescaled. The dashed grey line is a power law with exponent –1 as a guide for the eye. (**c**), Average number of clusters as a function of the total urbanised area, N_a, for the 128 × 128 km^2 United States regions (circles). The lower and upper values of the dashed areas denote the 10th and 90th percentile of 100 realisations of the null model. For clarity, curves have been shifted by m^2 along the x-axis. (**d**), Counter Cumulative Distribution Functions of populations of the clusters in all United States regions of 128 × 128 km^2 having urbanised area up to 5%. Regions are grouped in six groups according to their total population, N_x, and the CCDFs of each group are computed separately for the different values of the CCA parameter m. (**e**), The CCDFs of panel (**d**) collapse on the same curve when the axes are properly rescaled. The dashed grey line is a power law with exponent –1 as a guide for the eye. (**f**), Average number of clusters as a function of the total population, N_x, for the 128 × 128 km^2 United States regions (circles). The lower and upper values of the dashed areas denote the 10th and 90th percentile of 100 realisations of the null model. For clarity, curves have been shifted by m^2 along the x-axis

Local distributions of areas and populations of urban clusters To understand how the number, areas and populations of clusters depend on the CCA parameter m, we perform a systematic analysis of the GRUMP data, considering regions at a smaller spatial scale. Our first result is that the assumptions of the null model hold locally for small regions of size 128 × 128 km^2: the local distributions of city sizes are power laws with the same exponent as the global distribution at the country scale, with cutoffs that account for the finite sizes of the regions.

We divide the United States into non-overlapping square regions of size $L = 128$ km and obtain the urban clusters in each region applying the CCA for all values of m between 10 and 600. We group together regions with similar total population, N_x, and built-up area, N_a, and compute the distributions of cluster sizes (i.e. populations and areas) separately for each each group (see Fig. 5(a), (d)). In order to avoid finite-size effects, we only consider regions with low urbanisation, having a percentage of built-up area smaller than 5% (this condition is satisfied by 49% of the regions for $m = 10$ and up to 93% for $m = 599$). If the assumptions of our null model hold, the Counter Cumulative Density Functions (CCDFs) of cluster areas and populations should be truncated power laws and have the forms $P(> a|N_a) \sim a^{-\alpha_a} f_a(a/N_a)$ and $P(> x|N_x) \sim (x/m)^{-\alpha_x(m)} f_x(x/N_x)$, where f_a and f_x are scaling functions that rapidly go to zero when their argument is larger than 1, to account for finite-size effects. The scaling collapses shown in Fig. 5(b), (e) provide a validation to the predicted functional forms of the CCDFs.

Heap's law for urban clusters Our second result is that the average number of clusters is related to the total size of the region as predicted by the null model and Equation (1). This means that cities are randomly distributed among the regions, even at small spatial scales.

For each group of regions with similar total population, N_x, and built-up area, N_a, we compute the average number of clusters for all values of the CCA parameter m, $\langle C|N_a, m\rangle$ and $\langle C|N_x, m\rangle$, and we check if these empirical values are compatible with the estimates of the null model. To this end, we draw (with replacement) city areas and populations from the respective empirical distributions, until given target total values, N_a and N_x, are reached. We repeat this procedure 100 times for increasing values of N_a and N_x. For each value of total area and population, N_a and N_x, we compute the mean number of cities obtained for those total values, $\langle C|N_a\rangle$ and $\langle C|N_x\rangle$, and the confidence intervals defined as the 10th and 90th percentiles of the number of clusters obtained in the 100 realisations. We observe that the empirical estimates of the average number of clusters lie within the null model's confidence intervals (see Fig. 5(c), (f)), confirming that empirical data is compatible with a random distribution of clusters within the regions.

5 Conclusion

We empirically verified that a null model of urbanisation where cities are randomly distributed in space produces correct estimates of the expected number of cities in regions of various sizes worldwide. This fact does not mean that cities are non-interacting and independent of each other. On the contrary, it is apparent that urban systems are strongly interacting [35]: internal migrations, for example, create a dependency in the dynamics of the population in various cities, with some cities increasing in size because others are decreasing. However, our analysis demonstrates that such interactions do not produce urbanisation patterns characterised by significant spatial correlations. It is important to highlight that this result has been obtained for regions of 128×128 km^2 in the United States, and further analysis on global urbanisation patterns at higher spatial resolution is needed to test the validity of this conclusion in other countries and at smaller spatial scales. Moreover, this result is expected to hold in regions where urbanisation is not too high. In the analysis of gridded population data we only consider regions with low urbanisation, having a percentage of built-up area smaller than 5%. This is done because in highly urbanised areas deviations from Zipf's law and Heaps' law are inevitable. In fact, in regions with large population density, urban clusters start to merge and, as a result, when population keeps increasing the number of clusters decreases instead of increases. Also, the distribution of cluster sizes loses its characteristic power law tail because of the emergence of one giant cluster. The characterisation of urban patterns in the regime of large population density requires the development of a different theoretical framework, which is a task left for future work.

Many empirical lists of cities, including the one we consider in this study, may suffer from issues that can lead to inaccurate estimates of the number of cities in a country because, for example, some cities may appear multiple times with different names whereas other cities may be missing. To understand the impact of these issues, we performed numerical tests where we randomly duplicated and removed cities and verified that the result of the scaling relationship between number of cities and total population is not affected.

The theoretical result relating the average number of cities to the total population, Equation (1), is completely general and applicable to all systems characterised by Zipf's law for the distribution of group sizes, including word counts, the size of biological genera, the number of employees in firms and views/popularity of Youtube videos. Equation (1) is particularly useful in the analysis of finite-size systems in order to account for deviations from Heaps' law.

Recently, Zipf's law has been shown to be connected to Taylor's law, which describes the scaling between fluctuations in the size of a population and its mean [36]. This suggests the presence of a general connection between the three laws, Zipf's, Taylor's and Heaps' laws. Further research is needed to determine how the three laws can emerge from processes for the evolution of city sizes that incorporate birth, death and migration events.

Additional material

Additional file 1: Supplementary information (PDF 1.8 MB)

Acknowledgements
We thank J Grilli for insightful discussions.

Funding
FS is supported by EPSRC grant EP/P012906/1.

Abbreviations
GRUMP, Global Rural-Urban Mapping Project.

Availability of data and materials
The *Geonames* dataset analysed during the current study is available in the The GeoNames geographical database repository, http://download.geonames.org/export/dump/. The *GRUMP* dataset analysed during the current study is available in the Socioeconomic Data and Applications Center (SEDAC) repository, http://sedac.ciesin.columbia.edu/data/set/grump-v1-population-count.

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
FS and CJ analysed data and wrote the paper. All authors read and approved the final manuscript.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 17 November 2017 Accepted: 27 June 2019 Published online: 05 July 2019

References

1. United Nations (1997) Principles and recommendations for population and housing censuses
2. Seto KC, Güneralp B, Hutyra LR (2012) Global forecasts of urban expansion to 2030 and direct impacts on biodiversity and carbon pools. Proc Natl Acad Sci 109(40):16083–16088
3. McNicoll G (2005) United Nations, Department of Economic and Social Affairs: world economic and social survey 2004: international migration. Popul Dev Rev 31(1):183–185
4. Angel S, Parent J, Civco DL, Blei A, Potere D (2011) The dimensions of global urban expansion: estimates and projections for all countries, 2000–2050. Prog Plann 75(2):53–107
5. Auerbach F (1913) Das gesetz der bevölkerungskonzentration. Petermanns Geogr Mitt 59:74–76
6. Zipf GK (1935) The psycho-biology of language. Houghton, Mifflin, Oxford
7. Ioannides YM, Overman HG (2003) Zipf's law for cities: an empirical examination. Reg Sci Urban Econ 33(2):127–137
8. Rybski D, Ros AGC, Kropp JP (2013) Distance-weighted city growth. Phys Rev E 87(4):042114
9. Makse HA, Havlin S, Stanley H (1995) Modelling urban growth. Nature 377(1912):779–782
10. Rozenfeld HD, Rybski D, Andrade JS, Batty M, Stanley HE, Makse HA (2008) Laws of population growth. Proc Natl Acad Sci 105(48):18702–18707
11. Frasco GF, Sun J, Rozenfeld HD, Ben-Avraham D (2014) Spatially distributed social complex networks. Phys Rev X 4(1):011008
12. Zanette DH, Manrubia SC (1997) Role of intermittency in urban development: a model of large-scale city formation. Phys Rev Lett 79(3):523

13. Leyvraz F, Redner S (2002) Scaling theory for migration-driven aggregate growth. Phys Rev Lett 88(6):068301
14. Marsili M, Zhang YC (1998) Interacting individuals leading to Zipf's law. Phys Rev Lett 80(12):2741
15. Gabaix X (1999) Zipf's law for cities: an explanation. Q J Econ 114(3):739–767
16. Eeckhout J (2004) Gibrat's law for (all) cities. Am Econ Rev 94(5):1429–1451
17. Gabaix X, Ioannides YM (2004) The evolution of city size distributions. In: Handbook of regional and urban economics, vol 4, pp 2341–2378
18. Gibrat R (1931) Les inégalités économiques. Recueil Sirey, France
19. Hernando A, Hernando R, Plastino A, Zambrano E (2015) Memory-endowed US cities and their demographic interactions. J R Soc Interface 12(102):20141185
20. Heaps HS (1978) Information retrieval: computational and theoretical aspects. Academic Press, San Diego
21. Gerlach M, Altmann EG (2014) Scaling laws and fluctuations in the statistics of word frequencies. New J Phys 16(11):113010
22. Lü L, Zhang ZK, Zhou T (2010) Zipf's law leads to Heaps' law: analyzing their relation in finite-size systems. PLoS ONE 5(12):e14139
23. Azaele S, Suweis S, Grilli J, Volkov I, Banavar JR, Maritan A (2016) Statistical mechanics of ecological systems: neutral theory and beyond. Rev Mod Phys 88(3):035003
24. Bettencourt LM (2013) The origins of scaling in cities. Science 340(6139):1438–1441
25. Arcaute E, Hatna E, Ferguson P, Youn H, Johansson A, Batty M (2015) Constructing cities, deconstructing scaling laws. J R Soc Interface 12(102):20140745
26. Vatant B, Wick M (2012) Geonames ontology
27. The World Bank, World Development Indicators (2010) Available from: http://data.worldbank.org/data
28. Christaller W (1933) Die zentralen Orte in Süddeutschland: eine ökonomisch-geographische Untersuchung über die Gesetzmässigkeit der Verbreitung und Entwicklung der Siedlungen mit städtischen Funktionen. University Microfilms
29. Lösch A (1940) Die räumliche Ordnung der Wirtschaft: eine Untersuchung über Standort. Wirtschaftsgebiete und internationalen Handel, Verlag Wirtschaft und Finanzen. Jena
30. Center for International Earth Science Information Network CUFPRIWBCIdAT (2007) Global rural urban mapping project (GRUMP) alpha: gridded population of the world, version 2, with urban reallocation (GPW-UR)
31. Ratcliffe M, Burd C, Holder K, Fields A (2016) Defining rural at the US Census Bureau. American Community Survey and Geography Brief
32. Batty M, Longley PA (1994) Fractal cities: a geometry of form and function. Academic Press, San Diego
33. Lemoy R, Caruso G (2018) Evidence for the homothetic scaling of urban forms. In: Environment and planning B: urban analytics and city science p 2399808318810532
34. Nordbeck S (1971) Urban allometric growth. Geogr Ann, Ser B, Hum Geogr 53(1):54–67
35. Hernando A, Hernando R, Plastino A (2013) Space-time correlations in urban sprawl. J R Soc Interface 11(91):20130930
36. James C, Azaele S, Maritan A, Simini F (2018) Zipf's and Taylor's laws. Phys Rev E 98(3):032408