PNAS Nexus, 2024, 3, pgae329
https://doi.org/10.1093/pnasnexus/pgae329
Advance access publication 9 August 2024
Research Report

Indication of long-range correlations governing city size

Yunfei Li [a,b], Deniz Ural [a], Jan W. Kantelhardt [c] and Diego Rybski [b,d,e,*]

[a]Urban Transformations, Potsdam Institute for Climate Impact Research – PIK, Member of Leibniz Association, Potsdam 14412, Germany
[b]Research Area Spatial Information and Modelling, Leibniz Institute of Ecological Urban and Regional Development (IOER), Dresden 01217, Germany
[c]Institute of Physics, Martin-Luther-University, Halle (Saale) 06120, Germany
[d]Urban Living Lab Center (ULLC) a UN-Habitat Collaborating Center, Wuppertal Institute for Climate, Environment and Energy, Wuppertal 42103, Germany
[e]Complexity Science Hub Vienna, Vienna A-1090, Austria
*To whom correspondence should be addressed: Email: ca-dr@rybski.de
Edited By Attila Szolnoki

Abstract

City systems are characterized by the functional organization of cities on a regional or country scale. While there is a relatively good empirical and theoretical understanding of city size distributions, insights about their spatial organization remain on a conceptual level. Here, we analyze empirically the correlations between the sizes of cities (in terms of area) across long distances. Therefore, we (i) define city clusters, (ii) obtain the neighborhood network from Voronoi cells, and (iii) apply a fluctuation analysis along all shortest paths. We find that most European countries exhibit long-range correlations but in several cases these are anti-correlations. In an analogous way, we study a model inspired by Central Places Theory and find that it leads to positive long-range correlations, unless there is strong additional spatial disorder—contrary to intuition. We conclude that the interactions between cities extend over large distances reaching the country scale. Our findings have policy relevance as urban development or decline can affect cities at a considerable distance.

Keywords: city size, long-range correlations, spatial network, regional development

Significance Statement

It is well known that the city sizes in a country or region span a wide range of scales, e.g. from thousands to millions. The same holds for the area covered by the cities. However, little is known about the location of cities. Are large cities found next to each other, are large cities surrounded by small ones, or are they overall positioned randomly? In this article, we study correlations between neighboring city areas, second neighbors, …, from one side of the country to the other. It turns out that there are so-called long-range correlations across the entire countries. This implies that (urban) development is not restricted to the respective region but can have influence far beyond.

Introduction

Cities and urban systems exhibit a range of intriguing statistical regularities. Many of them are represented by scaling laws, which due to scale-invariance and self-similarity are particularly interesting. Batty (2013) lists seven laws of scaling (1, p.38ff) and one of them, Auerbach–Lotka–Zipf (ALZ) Law (2), states that the distribution of city sizes within a country or region follows a power-law. However, the distribution says nothing about the location of the cities. In other words, different positioning of cities and settlements can have the same size distribution. Where the cities are located and how they are related to each other is a different property—size and position are complementary.

Little research has been dedicated to the spatial organization of *city systems*. For example, a regular spacing between cities and settlements has been reported (3). Other authors assume random locations of cities (4) or report indications of nonrandom location pattern (5). Spatial correlations have been found in the growth rates of population (6–8). An important concept regarding the organization of city systems is the Central Places Theory (CPT) introduced by Christaller (9) and extended by Lösch (10). According to CPT, the cities are organized in a hierarchical, hexagonal manner, such that cities of similar size or importance repulse each other (11). CPT is consistent with ALZ Law (12, 13), but its empirical validation is challenging (e.g. (14)). Apparently, city systems do not follow the *ideal* CPT. They are messy, but at the same time not completely random. The organization within this stochasticity cannot be easily identified. Moreover, higher-order effects, such as polycentric urban organization characterized by multiple centers that are both balanced and in proximity (e.g. (15, 16)), are not included in CPT.

Competing Interest: The authors declare no competing interest.
Received: April 8, 2024. **Accepted:** July 29, 2024

We address this contrast by studying spatial correlations in the logarithmic sizes of cities and settlements across scales. A polycentric organization is characterized by cities and settlements of similar size in proximity, e.g. pronounced regions with large agglomerations and others with small ones. Contrary to this, cities can also have alternating sizes, where a small settlement is next to a large one, and a less developed region is neighboring an urbanized one. Studying land-cover data of European countries, we indeed identify long-range correlations (see Methods for a comparison and discussion of short-range and long-range correlations). They are a consequence of interactions among city sizes extending over a large range of scales. However, we find countries with positive and countries with negative correlations, but countries with positive correlations are more frequent, indicating a polycentric organization. Positive long-range correlations are consistent with simulations conducted with a simple model inspired by CPT. Our results suggest that interventions in one city can affect distant cities, and distant cities can also influence the effectiveness of such interventions. Therefore, it contributes to a better understanding of spatial correlations in city sizes which is crucial for developing effective regional and urban policies.

Data and methods

Data sources

CORINE land cover data (CLC) from Copernicus Land Monitoring Service (https://land.copernicus.eu/) for the year 2018 (version v.2020_20u1) represents the main data source. It comes in 100 m resolution and in GeoTiff format. The coordinate reference system is the standard European Coordinate Reference System defined by the European Terrestrial Reference System 1989 (ETRS89) datum and Lambert Azimuthal Equal Area (LAEA) projection (EPSG: 3035). The CLC2018 database covers the European area of EEA38 countries and the United Kingdom, see CORINE Land Cover User Manual (17, p.25). The standard CLC nomenclature includes 44 land cover classes, grouped in a three-level hierarchy. Five main categories are artificial surfaces, agricultural areas, forest and semi-natural areas, wetlands, and water bodies. In our work, all cells belonging to the artificial surfaces category are aggregated to one urban class, everything else is considered nonurban.

We used NUTS (Nomenclature of territorial units for statistics) level 1 data of the year 2021 from EUROSTAT (https://ec.europa.eu/eurostat/) to delineate countries. The data come in vector format (spatial shape) and it covers 37 countries (United Kingdom + all European Environment Agency member countries except Kosovo and Bosnia and Herzegovina).

Network construction

We apply the City Clustering Algorithm (CCA) with distance threshold l in order to define city clusters (8, 18). In CCA, any two sites (pixels) i and j are assigned to the same cluster if their Euclidean distance is smaller or equal to the threshold, i.e. $l_{ij} \leq l$, analogous to Random Geometric Graphs (19). We use an R-implementation of CCA (20). With increasing l, at certain point l_c there is a percolation transition (21–23). In order to avoid a system-spanning cluster, we choose values $l \ll l_c$.

Islands and land masses that are separated from the mainland of their corresponding country with a distance larger than 1 km are excluded. For example, Sardinia Island (Italy) and Corsica Island (France) are removed because the sea represents a natural barrier that affects the neighborhood relationships of the cities and settlements.

Next, we generate Voronoi polygons around the urban clusters. The distance from any nonurban cell within the Voronoi polygon to this urban cluster is always smaller than the distance to other urban clusters. The concept is the same as Voronoi cells (also known as Thiessen polygons in the geographic sciences) for spatial points, only that the Voronoi polygons are created for clusters, which have a spatial extent, instead of for points. The algorithm loops over every nonurban cell to find its closest urban cluster and then allocates it to the Voronoi polygon corresponding to the nearest urban cluster. When in some rare cases, the nearest urban cluster of a nonurban cell is not unique, then this cell is randomly associated to the Voronoi polygon corresponding to one of the nearest urban clusters. We define two urban clusters as neighbors if their Voronoi polygons share a border (based on Moore neighborhood). Implausible shortcuts at the coasts are avoided by limiting the Voronoi polygons to the land masses (within the country border). Based on the neighborhood table of all the urban clusters, we construct an undirected, unweighted network (see Fig. 1). For some analysis, we also consider the distance between the clusters, which we define as the Euclidean distance between the centers of mass of the neighboring clusters.

Analysis of long-range correlations

Then, we apply Shortest Path Fluctuation Analysis (SFA) (24) to the networks to characterize the correlation structure of the logarithmic cluster sizes (log of pixel count), $\log m_i$, along all shortest paths between all pairs of cities. For one of the longest paths through Germany with $i = 1, \ldots, N = 170$, Fig. 2 illustrates the advantage of studying the fluctuation function $F(d)$ (see below) instead of studying the auto-correlation function $C(d) = \frac{1}{N-d}\sum_{i=1}^{N-d}(\log m_i \times \log m_{i+d})$. Figure 2A and B shows the real (long-range correlated) logarithmic cluster sizes $\log m_i$ and a (short-range correlated) realization of an AR1 process, respectively. In Fig. 2C, the corresponding auto-correlation functions are plotted. Due to the limited statistics with merely 170 data points, they fluctuate a lot, so that it is difficult to distinguish between the power-law decay for the real data, $C(d) \sim d^{-0.4}$ ($\gamma = 0.4$), and the exponential decay for the AR1 data, $C(d) = \exp(-d/0.97)$. The scaling behavior of the real data can be identified more easily by studying the fluctuation function $F(d) \sim d^{H-1} = d^{-\gamma/2}$ in Fig. 2D, since the long-range correlations of the real data are reflected in a nice scaling behavior with slope $-\gamma/2 = 1 - H = -0.2$ in the double-logarithmic plot, while the short-term correlations of the AR1 data are reflected in a line crossing over from an initially larger slope to $1 - H = -0.5$ in the limit of large d. We have also included the result for shuffled real data, where uncorrelated behavior is reflected by the $H = 0.5$ (i.e. slope −0.5) since all correlations have been destroyed by the shuffling.

In SFA, the standard deviation F of averages of values along shortest paths of length d is analyzed. The length is measured in terms of network steps. SFA consists of the following steps. (i) Find the shortest path between all pairs of nodes. (ii) Calculate the average of the logarithmic cluster sizes along the considered shortest path. (iii) For a fixed shortest path length d calculate the standard deviation of those averages. (iv) Plot this standard deviation as a function of path length, i.e. $F(d)$ vs. d. If the values associated to the nodes are long-range correlated, then $F(d) \cdot d \sim d^H$, where H is analogous to the Hurst-exponent, i.e. positive correlations are measured by $H > 0.5$ and negative ones by $H < 0.5$. If they are uncorrelated then $H \simeq 0.5$, i.e. $F(d) \sim d^{-0.5}$, is

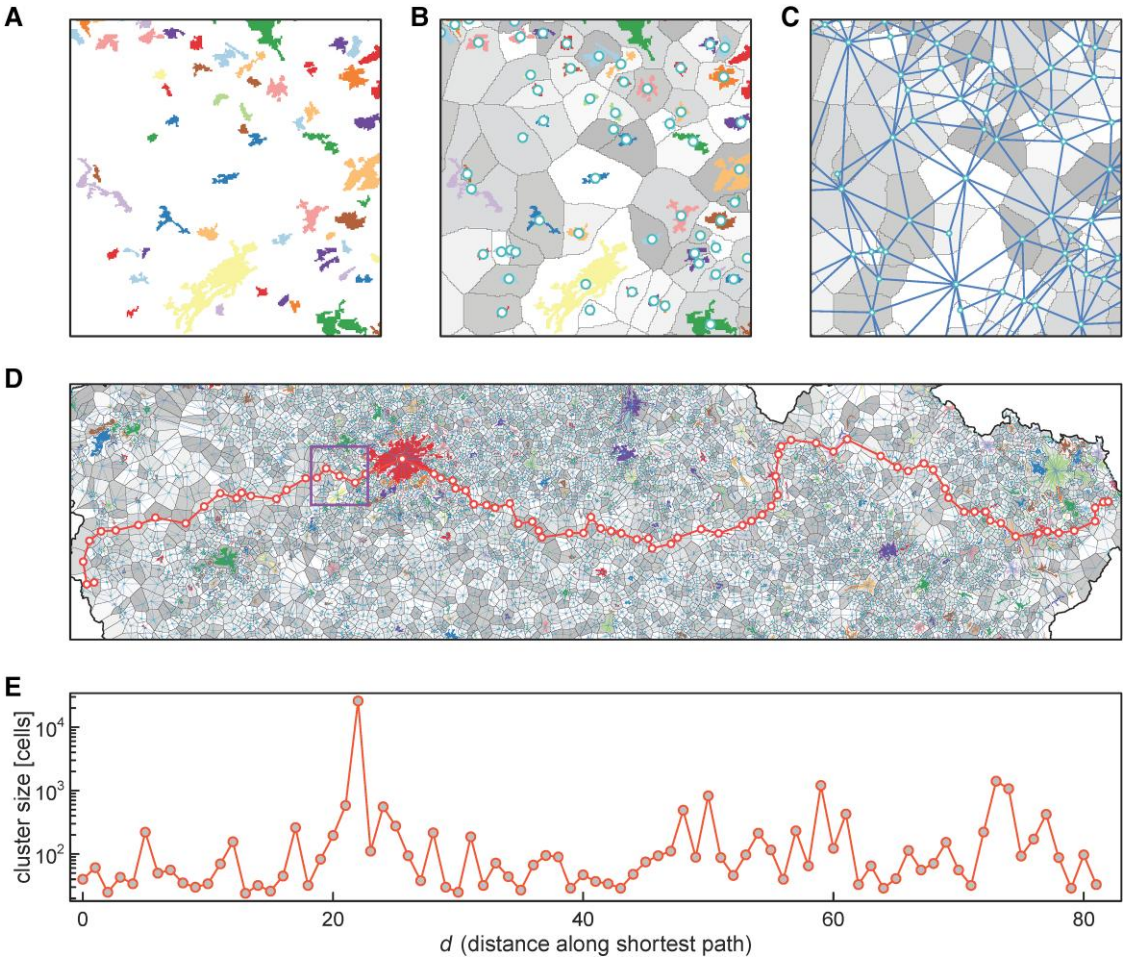

Fig. 1. Illustration of steps to obtain a sequence of city sizes in the Czech Republic. A) Spatial clustering of urban land cover. Urban pixels are clustered employing the City Clustering Algorithm (CCA) with a distance threshold of $l = 200$ m, i.e. any two urban sites belong to the same cluster if their distance is smaller than or equal to l. B) Voronoi polygons. Any nonurban site is associated to its closest urban cluster. This set of nonurban sites forms the Voronoi polygon of the respective urban cluster. Cells in different shades of gray represent the Voronoi polygons, circles indicate the center of mass of each urban cluster. C) Settlement network. Two clusters are considered as adjacent nodes (connected by one edge in solid line) if their corresponding Voronoi polygons touch each other. D) Shortest path. The shortest path between any two urban clusters is determined—here an example is highlighted in red. The purple rectangle on the map is the outer box of the area shown in the three top panels. E) Sequence of sizes. Along the shortest path (D) the respective cluster size (number of grid cells) is plotted on a logarithmic scale vs. the number of steps.

found. The exponent H is obtained from ordinary least squares regression to the log-quantities. In order to be able to use the same fitting range for various l and different countries, we rescale by plotting $F(d)$ as a function of d/D, where D is the length of the longest shortest path (diameter) of the considered network.

In order to assess the significance of the measured exponents H, we apply a shuffling. For the networks extracted from each country at each distance threshold, we shuffle the sizes associated to the nodes (i.e. randomization is done by shuffling the sizes associated with the nodes; this preserves the network structure and the size distribution) and then repeat SFA. This destroys correlations between the nodes and sizes but preserves all other properties including the network itself and the cluster size distribution. We repeat the shuffling 10 times for each network. For the estimated $H(l)$ at CCA distance threshold value l of a country, we compare $H(l)$ with the average value of the estimated H_s values from the 10 shuffling realizations. We test whether H is significantly larger (H_0: $H \leq \mu(H_s)$) or smaller (H_0: $H \geq \mu(H_s)$) than the H_s-values using the Z-test. Denoting the average of the H_s-values as $\bar{H}_s$, for H_0: $H \leq \mu(H_s)$, if $P(h > \frac{H-\bar{H}_s}{\sigma(H_s)}) \leq 0.001$ we reject the null hypothesis and consider H to be significantly larger than H_s. Similarly, for H_0: $H \geq \mu(H_s)$, if $P(h < \frac{H-\bar{H}_s}{\sigma(H_s)}) \leq 0.001$, we reject the null hypothesis and consider H to be significantly smaller than H_s.

Modeling central places

We employ a model that generates structures inspired by Christaller's Central Places Theory (CPT) (9, 25, 26) and that resembles the model proposed in (27). It starts (0th iteration) with a single point carrying the size $s^{-0+\mathcal{N}(0,0.5)}$, where s is a parameter that determines how the node size decreases with the iteration, and $\mathcal{N}(0, 0.5)$ represents a random number drawn from the normal distribution with $\mu = 0$, and $\sigma = 0.5$. At each iteration $i > 0$, 6 points are added hexagonally, at distance of 2^{-i}, around the points added in the $(i-1)$th realization and carry the size $s^{-i+\mathcal{N}(0,0.5)}$. Points at the same position are removed. After finishing the process at ith generation, the distances between the neighboring nodes are $\mu_L = 2^{-i}$. For any point $P(x, y)$, we add noise to its coordinates, $P'(x', y') = (x + \mathcal{N}(0, n_p\mu_L), y + \mathcal{N}(0, n_p\mu_L))$, where $\mathcal{N}(0, n_p\mu_L)$ represents a random number drawn from the normal distribution with $\mu = 0$, and $\sigma = n_p\mu_L$, n_p is another parameter (ranging from 1–300%) that controls the spatial disorder level of the structure.

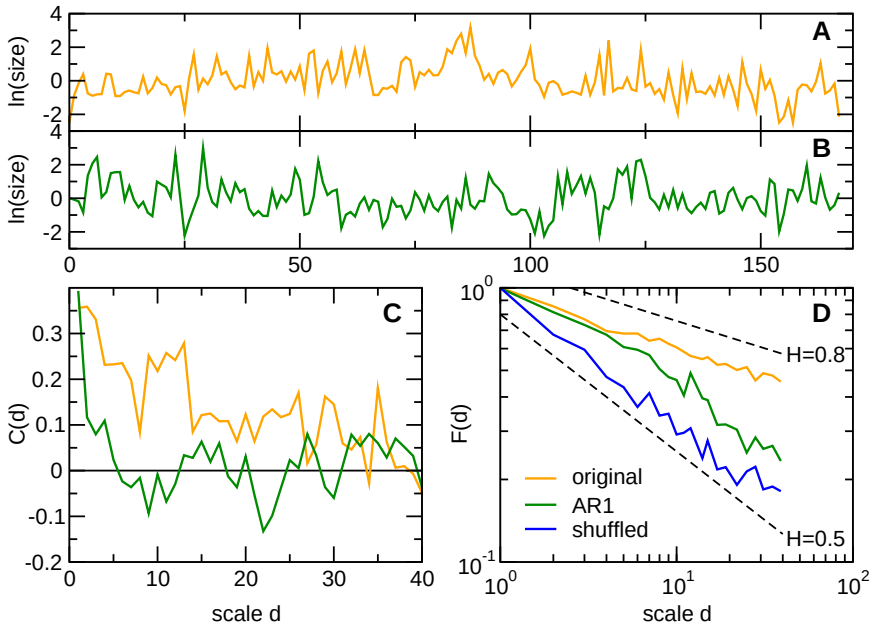

Fig. 2. Example to illustrate the advantage of fluctuation function over auto-correlation function. A) Logarithmic cluster sizes of one of the longest shortest paths for Germany and $l = 100$. B) Data generated with the auto-regressive model (AR1) using the same correlation length $C(d = 1) = 0.356$ as observed in (A). C) Auto-correlation functions for the records shown in (A) and (B), respectively. D) Fluctuation functions for the records shown in (A) and (B). An additional curve represents the fluctuation function of a shuffled version of the data from (A). The dashed lines serve as guide to the eye and have slopes −0.2 (for $H = 0.8$) and −0.5 (for $H = 0.5$).

For each output from the CPT model, we generate a Voronoi diagram based on the coordinates of all the nodes and build the unweighted and undirected network to connect the nodes when their Voronoi polygons touch each other. It has to be noted that, to avoid shortcuts between the nodes on the boundary, we clip the Voronoi diagram using the convex hull of all the points. This prevents two nodes from being considered as neighbors if their Voronoi polygons touch each other only outside the convex hull. Then, we apply SFA analogous to the real-world data. We use shortest paths in terms of Euclidean distance since for small spatial noise the hexagonal structure persists with a multitude of shortest paths based on network steps.

Results

Analyzing real-world data

We begin by constructing settlement networks. Since administrative units hamper the identification of neighborhood relations, we analyze urban land cover data and define urban clusters. This spatial clustering involves a distance threshold l which determines if two urban sites are part of the same cluster. Examples are shown in Fig. 1A. Two urban clusters are then considered neighbors, if their Voronoi polygons share a border. In Fig. 1B, corresponding Voronoi polygons are exemplified, and in Fig. 1C, the respective network is displayed, where (for visualization purpose) we use the centers of mass as node positions. This procedure resembles Delaunay triangulation and the neighbors of an urban cluster are uniquely defined while being dependent on the aggregation scale represented by l.

If we want to analyze correlations beyond nearest neighbors, we somehow need to consider the neighbors of the neighbors and so forth. An intuitive way of defining them is the shortest path on the network. Figure 1D shows an example of a long shortest path. The corresponding sequence of logarithmic cluster sizes (in terms of area) along this shortest path is provided in Fig. 1E. It can be studied analogous to a time series and long-range correlations come along with extended regions of in- or decreased values (28, 29). Specifically, positive long-range correlations are reflected in a power-law decay of the auto-correlation function $C(d) \sim d^{-\gamma}$ with distance d and $0 < \gamma < 1$, a power-law decay of the power spectrum, $P(f) \sim f^{-\beta}$ with frequency f and $\beta = 1 - \gamma$, and a fluctuation function $F(d) \sim d^{H-1}$ with $H = 1 - \gamma/2$ (Fig. 2 in Methods section). Such long-range correlations are due to interactions that extend across large spatial scales. They are clearly distinct from short-range correlations that would result, e.g. from an auto-regressive process.

The shortest path between two urban clusters could be a special case. Hence, in order to make best use of the statistics, we take into account the shortest path between all pairs of nodes which is part of SFA (24) (Methods section). Resulting fluctuation functions are depicted in Fig. 3 for $l = 200$ m. The examples, Bulgaria and Germany, clearly exhibit positive long-range correlations, their fluctuation functions decrease more slowly than in the uncorrelated case. The corresponding exponents $H > 0.5$ are also very different from those that we obtain for the shuffled data $H_s \approx 0.5$. We conclude that in these cases neighboring urban clusters have related sizes. The correlations among the clusters extend at least up to one-third of the size of the countries (i.e. the upper limit of our fitting range).

Interestingly, the other examples, Spain and Finland, exhibit fluctuation functions that decrease steeper than in the uncorrelated case for the shuffled data. The exponents $H < 0.5$ indicate long-range anti-correlations, meaning that neighboring urban clusters are alternating in size—again across a large range of spatial scales. For other European countries, we obtain similar results, not always significant but in most cases $H \neq 0.5$ (Fig. 3E). It is noticeable that not all coastal countries exhibit negative correlations but almost all countries exhibiting negative correlations have extended coasts (an exception is Austria; Slovenia has a short coast).

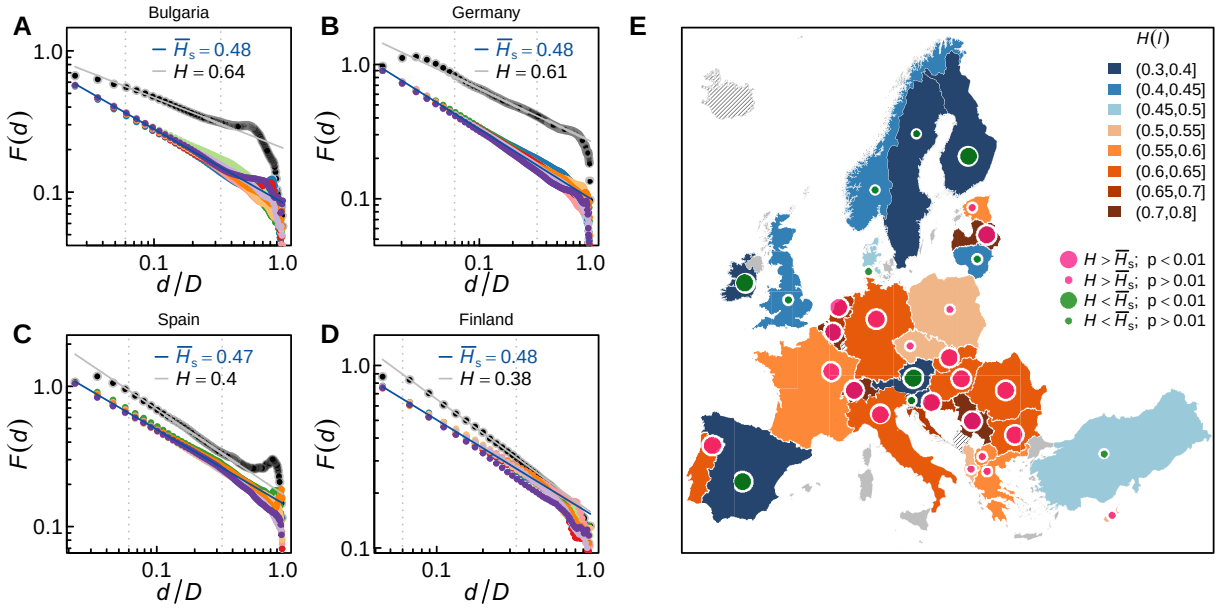

Fig. 3. Fluctuation analysis and resulting fluctuation exponents. A–D) show the fluctuation functions (for $l = 200$ m) of four example countries (Bulgaria, Germany, Spain, and Finland, respectively). The fluctuation functions F are plotted vs. distance d in terms of network steps divided by network diameter D. The circles represent the obtained values and the solid line indicates the estimated slope H. The colored symbols stem from analyses where the association between nodes and cluster size is destroyed by shuffling (10 realizations). The respective slopes H_s are used to infer the significance of the original results. $\bar{H}_s$ is the average of the respective H_s-values. The respective solid line indicates the slope of all shuffled realizations. For both real-world network and shuffle exercises, only the range of $d/D \in [0.06, 0.33]$ (delineated by the vertical dashed lines) is used to estimate the exponents. E) Map of estimated slopes H at $l = 200$ m. The color within the country borders indicates the obtained slopes. Hatched areas represent countries with fewer than 5 points falling within the chosen fitting range or with insignificant ($p > 0.001$) regression, gray areas are distant land masses that are at least 1 km apart from the major land mass of their corresponding counties. The pink circles on top of each country indicate that H is larger than the average of the H_s from shuffle exercises, while the size of them informs if the Z-test indicates significance or not, analogous for the green circles but for $H < H_s$. Many countries in central Europe exhibit significant positive long-range correlations, while several others also exhibit negative correlations.

In Fig. 4, resulting fluctuation functions and exponents are shown for various clustering thresholds l. The respective fluctuation functions (Fig. 4A–D) widely agree with those of Fig. 3. We see no systematic dependence when plotting the fluctuation exponents vs. the clustering threshold in Fig. 3E. There are some variations but the exponents to a large extent remain different from $H = 0.5$ and consistently above or below this limit (except for Czech Republic). Increasing the aggregation scale with l corresponds to a coarse-graining (Fig. S1) and is comparable to aggregating a monthly time series into an annual one. Since power-law correlations are scale-invariant, the correlations should not be affected by aggregation. This is also the case for the correlations among city sizes (Fig. 3E) and supports the robustness of our results.

Alternative shortest path

Judging from Fig. 1D one may object that the shortest path follows a somewhat curvy and arbitrary route (30). It is also evident that the shortest path in terms of network steps gives preference to large Voronoi polygons, which can belong to large urban clusters but also to small ones in remote areas. Accordingly, we repeat the analysis employing the shortest path in terms of Euclidean distance, i.e. the cumulative distance between the urban clusters is minimized. The respective results are shown in Fig. S2. The shortest path defined by Euclidean distance is very different from the one shown in Fig. 1D. But when we compare the estimated exponents from both methods, we obtain a correlation coefficient of 0.46. Separating the values by distance threshold l we find correlations beyond 0.6 for $l = 200$ m and $l = 300$ m and below 0.4 for $l = 100$ m and $l \geq 700$ m. The shortest path in terms of Euclidean distance is also not ideal as it avoids large Voronoi polygons. Nevertheless, the comparison of both variants shows that the results are similar, indicating that the influence of the actual path is marginal.

Modeling long-range correlations in city size

Last, we want to employ a numerical model to test under which circumstances long-range (anti-) correlations can emerge in city size. Inspired by Central Places Theory (9, 27) we generate structures consisting of points and associated city sizes. In a self-similar manner, six smaller cities surround a respective larger one—hierarchically over a range of size scales (Fig. 5A). In order to be more similar to real-world structures, we add noise to the positions and to the sizes (Fig. 5B). We then analyze them analogous to the real-world data using shortest paths in terms of Euclidean distance (see Data and methods). In Fig. 5C, we plot the resulting fluctuation exponents as a function of the magnitude of spatial noise (see Fig. S3 for examples of generated structures and resulting fluctuation functions). One can see that for a small degree of noise we obtain positive long-range correlations which then with increasing noise transition to (weak) negative ones and asymptotically vanish completely. From these simulations, we learn that the model inspired by CPT can lead to long-range correlations. Whether the magnitude of spatial noise is also the driving factor of long-range correlations in the real-world data is a complex problem and needs to remain for future research.

Discussion

In summary, we quantify long-range correlations in logarithmic city size by combining spatial clustering, Delaunay triangulation,

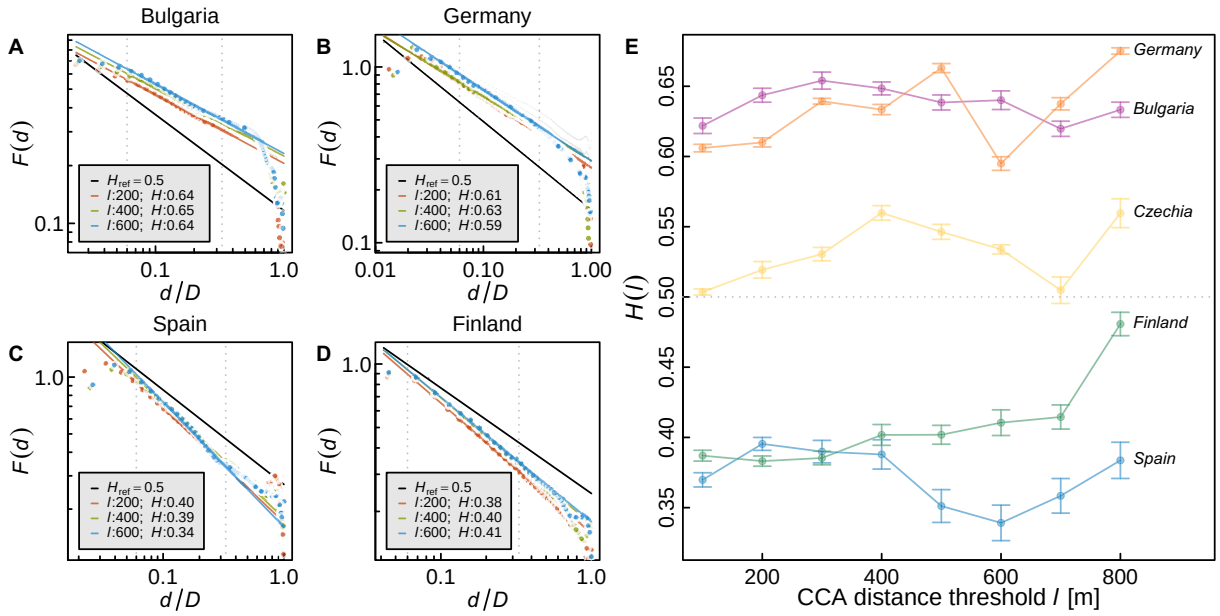

Fig. 4. Fluctuation exponent dependence on the clustering threshold l. Analogous to Fig. 3, A–D) show the fluctuation functions for Bulgaria, Germany, Spain, and Finland but here the curves for clustering thresholds $l = 400$ m and $l = 600$ m are also included. Colored solid lines represent respective slopes from regression; for comparison the solid lines indicates the uncorrelated case $H = 0.5$. Asterisks in the background represent the values from all $l \in \{100, 200, \ldots, 800\}$ m. E) shows the estimated $H(l)$ against CCA distance threshold l, error bars denote the standard deviation of the estimated values by fitting the fluctuation function in the range $d/D \in [0.06, 0.33]$. Here, additionally the results for Czech Republic are included which indicate minor or absence of correlations. Overall, the influence of l is marginal and the correlations are mostly scale-independent.

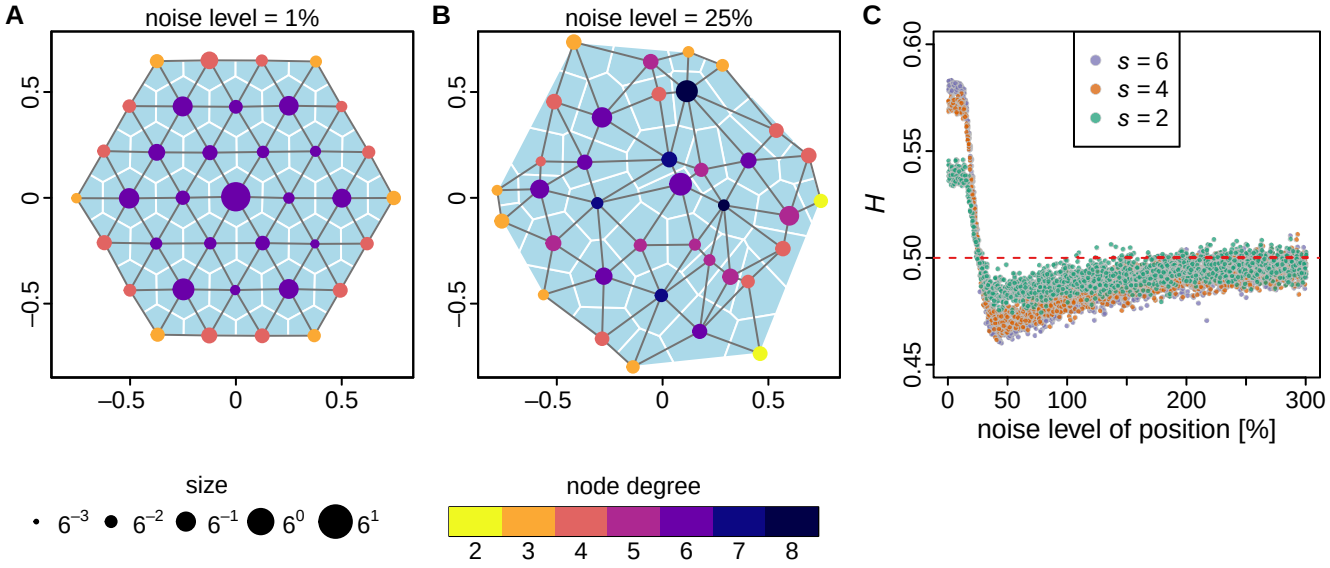

Fig. 5. Simulations of structures inspired by Central Places Theory (CPT) and dependency of the fluctuation exponent on the level of spatial noise. In each model iteration, six smaller settlements are located around those of the previous iteration, whereas the size, associated with each settlement, decreases from iteration to iteration by a factor s. Gaussian noise is multiplied and added to the resulting sizes and positions, respectively. A) Small example created with the model with 3 iterations (37 points) and 1% spatial noise. The areas represent the Voronoi polygons of each node, the nodes are considered as neighbors (connected by solid edges) if their Voronoi polygons touch each other. To minimize the boundary effect, the Voronoi polygons do not extend beyond the convex hull of the structure. The size and color indicate the size and degree of each node, respectively. The degree is included to illustrate how the noise affects the regular structure. B) As before but with 25% spatial noise. The small structures are just for illustration purpose, in Fig. S3D, E two larger CPT structures with different noise levels are visualized, with some small-scale details zoomed in Fig. S3F, G. C) Dependence of H on the level of spatial noise in the CPT model output with six iterations (12,097 nodes, see Fig. S3 for some examples of the structures and resulting fluctuation functions). Different values of the factor s change the resulting fluctuation exponents but do not affect the overall picture. Depending on the level of spatial noise, positive as well as negative long-range correlations can be found in the structures generated by the model.

and SFA (originating from complex networks theory). We obtain exponents that significantly differ from the uncorrelated case but whether above or below is specific to the considered countries. Countries with positive long-range correlations are more frequent in Europe suggesting extended regions with large or small cities. We note that we have obtained similar, but less stable results for

studying city sizes instead of logarithmic city sizes, mainly because of the extreme (exponential) variations in the actual city sizes (also present in the CPT model). The simulations with the CPT-inspired model result in positive long-range correlations for structures close to hexagonal organization. This seems counter-intuitive, because the CPT model is based on inserting small cities around each large city in each generation.

Long-range correlations in city size are associated with interactions between the cities and settlements that extend over many spatial scales. Certainly, a large city influences its vicinity. But the long-range character of the interactions—as revealed by our results—indicates that the influence extends far beyond the neighboring cities and settlements. This means, the size of a city or settlement is a result of the sizes of many other units, even far away. It can be assumed that such interactions not only affect city size (here we consider the area)—but also other city features, including socio-economic properties (e.g. (31, 32)). Accordingly, we also anticipate an influence on regional development and argue that such phenomena need to be taken into account by respective policies.

It stands out that countries with negative correlations are almost exclusively countries with long coasts (not vice versa, though). We speculate that the reason is coastal development. The coasts represent approximately linear objects along which coastal development happens with alternating small and large cities. Since many shortest paths follow this structure, the resulting fluctuation function exhibits negative correlations.

In continental countries, development happens potentially everywhere and many shortest paths rather consist of consecutive cities and settlements of similar size. This is best illustrated by simulations with our CPT-inspired model. Only distinct shortest paths exhibit the characteristic alternating structure. Apparently, many paths do not follow those distinct patterns but rather consist of extended segments of small or large cities. As a consequence, the model results are dominated by paths that do not follow distinct (anti-correlated) patterns but rather sequences of correlated sizes —which ultimately dominate the resulting fluctuation functions. Thus, the six units placed around the central one, having similar size, cause this counter-intuitive result.

Positive long-range correlations could also be indicative of development axes. Development axes are characterized by (growth) centers that align along an axis at a large (continental, country, or regional) scale (e.g. (33–35)). As such, development axes could be understood as a special form of polycentrism and are not explained by CPT. They represent more or less developed areas extending spatially and should manifest themselves in spatial correlations. Consequently, the long-range correlations in city size that we measure here could be associated with development axes.

Although in itself consistent, our work also leaves room for improvement. When two settlements are separated by a mountain range, they might not be perceived as neighbors. Accordingly, including further data, such as topographic information, might enable a refined neighborhood relation of cities and settlements. Similarly, one could include road-network data to define neighbors and networks (e.g. (36)). However, as our results are somewhat robust against the choice of the shortest path, we do not expect a dramatic influence from such refinements. For the sake of simplicity and consistency, we base the entire analysis on one and the same data set (land-cover).

Regarding the simulations with the CPT model, we need to note that an important feature is missing, namely the spatial extent of the cities. In our simulations, all nodes have the same spatial size and only carry an additional attribute representing the city size. We cannot exclude that improved simulations, taking this factor into account, may lead to different results. However, it is plausible that, qualitatively, both versions should lead to similar outcomes.

There are many directions in which our work can be extended in future research. While there are countries with positive and negative long-range correlations, some countries also exhibit an absence of correlations in city size. Although the precise mechanisms behind our empirical observations require further investigation, it seems plausible that this absence might reflect a simultaneous presence of effects leading to positive and negative correlations, compensating each other.

Further work will be necessary to understand the influence of historic borders and planned cities. It is plausible that historical and political developments affect current city configurations, e.g. Eastern Germany, Austria, and the former Yugoslavia. Follow-up work could address historic and country-specific questions, e.g. involving sub-national analyses.

Moreover, examining urban land-cover dynamics over time can enhance our understanding of how the long-range spatial correlation of city sizes impacts urban development patterns. This might be achieved by systematically exploring the spatio-temporal relationship between urban growth rates and the long-range correlation of city sizes.

Certainly, more conceptual work is necessary to discriminate and understand polycentrism in the context of Central Places Theory. Similarly, the role of inhomogeneities needs to be better understood and related to polycentrism. It is very common that countries exhibit more or less densely populated regions. To what extent do such changes represent noise or fluctuations, and to what extent are these systematic nonstationarities?

Methodologically, instead of Delaunay Triangulation one could also employ other proximity graphs (37), including Euclidean relative-neighborhood graphs (e.g. (38, 39)) or the Gabriel Graph (e.g. (40)). A more pragmatic alternative could be to define the network by connecting neighboring cities if they lie within a predefined Euclidean distance (41). Instead of SFA one could explore fractal network analysis (42, 43), likely to provide complementary exponents. Last but not least, it should be straightforward to expand the analysis to other countries and continents.

An alternative simulation model could be based on the Dodds network discussed by Aste (44). The idea is to cover a continuous surface with nonoverlapping circles. After placing circles with a constant maximum radius R_{max}, smaller circles (always chosen as large as possible) fill in the gaps. The spatial organization of this system resembles urban systems (3). They also share power-law size distributions. The Dodds network is then given by connecting the centers of neighboring circles and can be analyzed analogously to our Delaunay triangulation. It could also be insightful to study more sophisticated models, like SIMPOP (45) which is a multiagent system that can simulate characteristics of urban systems instead of imposing them as in our CPT implementation.

In the present work we study the correlations in city size. Alternatively, one could also investigate the degree correlations, for which SFA was developed originally. The degree is the number of connections a node has to others. In our context this would be the number of neighbors a city has. Then one could also study the analog to Aboav's law. For planar Poisson–Voronoi tessellations, Hilhorst (46) finds that the neighbors of a node with degree k have on average degree $k' = 4 + 3\sqrt{\pi/k}$. As cities are not Poissonian an exponent different from $-1/2$ can be anticipated in this relation.

Acknowledgments

We appreciate useful discussions with C. Rozenblat. We would also like to thank the reviewers for their insightful comments and valuable suggestions, which significantly improved the quality and clarity of this manuscript.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

Y.L. and D.R. thank the German Research Foundation (DFG) for funding this research within the *Urban Percolations* project (451083179). D.R. thanks the Alexander von Humboldt Foundation for financial support under the Feodor Lynen Fellowship. This work emerged from ideas discussed at the symposium *Cities as Complex Systems* (Hanover, 2016 July 13th–15th) which was generously founded by VolkswagenFoundation.

Author Contributions

Conceptualization: D.R.; Methodology: Y.L., D.U., J.W.K., and D.R.; Investigation: Y.L., J.W.K., and D.R.; Visualization: Y.L.; Supervision: J.W.K. and D.R.; Writing–original draft: D.R.; Writing–review & editing: Y.L., J.W.K., and D.R.

Data Availability

The extracted and analyzed network data, and the code for shortest path fluctuation analysis, are shared on Zenodo and can be publicly accessed at https://doi.org/10.5281/zenodo.11501688.

References

1 Batty M. 2013. *The new science of cities*. Cambridge, London: MIT Press.
2 Rybski D, Ciccone A. 2023. Auerbach, Lotka, Zipf – pioneers of power-law city-size distributions. *Arch Hist Exact Sci*. 77:601–613.
3 Glass L, Tobler WR. 1971. Uniform distribution of objects in a homogeneous field: cities on a plain. *Nature*. 233(5314):67–68.
4 Simini F, James C. 2019. Testing heaps' law for cities using administrative and gridded population data sets. *EPJ Data Sci*. 8:24.
5 González-Val R. 2019. The spatial distribution of US cities. *Cities*. 91:157–164.
6 Hernando A, Hernando R, Plastino A. 2014. Space–time correlations in urban sprawl. *J R Soc Interface*. 11(91):20130930.
7 Hernando A, Hernando R, Plastino A, Zambrano E. 2015. Memory-endowed US cities and their demographic interactions. *J R Soc Interface*. 12(102):20141185.
8 Rozenfeld HD, *et al.* 2008. Laws of population growth. *Proc Natl Acad Sci U S A*. 105(48):18702–18707.
9 Christaller W. 1966. *Central places in Southern Germany*. London: Prentice-Hall International, Inc. Translated from "Die zentralen Orte in Süddeutschland" by C. W. Baskin.
10 Lösch A. 1954. *The economics of location*. New Haven: Yale University Press. Translated from "Die räumliche Ordnung der Wirtschaft" [2nd edition, 1944; 1st edition, 1940] by W. H. Woglom [with the assistance of W. F. Stolper].
11 Mori T, Smith TE, Hsu W-T. 2020. Common power laws for cities and spatial fractal structures. *Proc Natl Acad Sci U S A*. 117(12):6469–6475.
12 Berry BJL, Garrison WL. 1958. Alternate explanations of urban rank-size relationships. *Ann Assoc Am Geogr*. 48(1):83–90.
13 Hsu W-T. 2012. Central place theory and city size distribution. *Econ J*. 122(563):903–932.
14 Shi L, Wurm M, Huang X, Zhong T, Taubenböck H. 2020. Measuring the spatial hierarchical urban system in China in reference to the central place theory. *Habitat Int*. 105:102264.
15 Derudder B, Meijers E, Harrison J, Hoyler M, Liu X. 2022. Polycentric urban regions: conceptualization, identification and implications. *Reg Stud*. 56(1):1–6.
16 Lemoy R. 2024. Monocentric or polycentric city: an empirical perspective. In: Rybski D, editor. Compendium of urban complexity. Springer, in preparation.
17 Büttner G, *et al.* 2021. Copernicus land monitoring service-corine land cover. User manual. Technical report, Copernicus Publications.
18 Rozenfeld HD, Rybski D, Gabaix X, Makse HA. 2011. The area and population of cities: new insights from a different perspective on cities. *Am Econ Rev*. 101(5):2205–2225.
19 Dall J, Christensen M. 2002. Random geometric graphs. *Phys Rev E*. 66(1):016121.
20 Kriewald S, Fluschnik T, Reusser D, Rybski D. 2019. *osc: orthodromic spatial clustering*. R package version 1.0.5.
21 Behnisch M, Schorcht M, Kriewald S, Rybski D. 2019. Settlement percolation: a study of building connectivity and poles of inaccessibility. *Landscape Urban Plan*. 191:103631.
22 Fluschnik T, *et al.* 2016. The size distribution, scaling properties and spatial organization of urban clusters: a global and regional percolation perspective. *Int J Geo-Information*. 5(7):110.
23 Hemond O, Rybski D, Wartenberg AC, Butsic V. 2023. Assessing fire landscape connectivity in California using predictive percolation. in preparation.
24 Rybski D, Rozenfeld HD, Kropp JP. 2010. Quantifying long-range correlations in complex networks beyond nearest neighbors. *EPL*. 90(2):28002.
25 Mulligan GF, Partridge MD, Carruthers JI. 2012. Central place theory and its reemergence in regional science. *Ann Reg Sci*. 48:405–431.
26 Ullman E. 1941. A theory of location for cities. *Am J Sociol*. 46(6):853–864.
27 Openshaw S, Veneris Y. 2003. Numerical experiments with central place theory and spatial interaction modelling. *Env Plan A*. 35(8):1389–1403.
28 Kantelhardt JW, Koscielny-Bunde E, Rego HHA, Havlin S, Bunde A. 2001. Detecting long-range correlations with detrended fluctuation analysis. *Physica A*. 295(3–4):441–454.
29 Peng C-K, *et al.* 1992. Long-range correlations in nucleotide sequences. *Nature*. 356(6365):168–170.
30 Kartun-Giles AP, Barthelemy M, Dettmann CP. 2019. Shape of shortest paths in random spatial networks. *Phys Rev E*. 100(3):032315.
31 Ribeiro FL, Rybski D. 2023. Mathematical models to explain the origin of urban scaling laws. *Phys Rep*. 1012:1–39.
32 Ribeiro HV, Oehlers M, Moreno-Monroy AI, Kropp JP, Rybski D. 2021. Association between population distribution and urban gdp scaling. *PLoS One*. 16(1):e0245771.
33 Güßefeldt J. 1978. Die graphentheorie als instrument zur beurteilung raumordnungspolitischer konzepte. Dargestellt am beispiel der entwicklungsachsen von baden-Württemberg und bayern. *Geogr Z*. 66(2):81–105.
34 Netrdová P, Nosek V. 2016. Spatial patterns of unemployment in Central Europe: emerging development axes beyond the blue banana. *J Maps*. 12(4):701–706.

35 Purboyo H, Santoso EB, Sawitri D. 2012. The development of local nodes along transportation corridors: a review of development axes theory. In: The 11th IRSA International Conference Proceeding. Banjarmasin: IRSA 11th committee.
36 Prieto-Curiel R, Schumann A, Heo I, Heinrigs P. 2022. Detecting cities with high intermediacy in the African urban network. *Comp Environ Urban Sys*. 98:101869.
37 Cimikowski RJ. 1992. Properties of some Euclidean proximity graphs. *Patt Recogn Lett*. 13(6):417–423.
38 Jaromczyk JW, Toussaint GT. 1992. Relative neighborhood graphs and their relatives. *Proc IEEE*. 80(9):1502–1517.
39 Melchert O. 2013. Percolation thresholds on planar Euclidean relative-neighborhood graphs. *Phys Rev E*. 87(4):042106.
40 Norrenbrock C. 2016. Percolation threshold on planar Euclidean Gabriel graphs. *Eur Phys J B*. 89:1–6.
41 Esch T, *et al*. 2014. Dimensioning urbanization – an advanced procedure for characterizing human settlement properties and patterns using spatial network analysis. *Appl Geogr*. 55:212–228.
42 Gallos LK, Song C, Havlin S, Makse HA. 2007. Scaling theory of transport in complex biological networks. *Proc Natl Acad Sci U S A*. 104(19):7746–7751.
43 Song C, Havlin S, Makse HA. 2005. Self-similarity of complex networks. *Nature*. 433(7024):392–395.
44 Aste T. 1996. Circle, sphere, and drop packings. *Phys Rev E*. 53(3): 2571–2579.
45 Sanders L, Pumain D, Mathian H, Guérin-Pace F, Bura S. 1997. SIMPOP: a multiagent system for the study of urbanism. *Environ Plan B*. 24(2):287–305.
46 Hilhorst HJ. 2008. Statistical properties of planar Voronoi tessellations. *Eur Phys J B*. 64:437–441.