Supplementary Materials

Abbreviations: E&W, England and Wales; CAS, Census Area Statistics; ST, Standard Table; SIC, Standard Industrial Classification

Unit of Geography

The underlying spatial unit for all city cluster aggregations is the Census Area Statistics (CAS) ward definition produced by the UK Office for National Statistics. Ward boundaries reflect the political geography of the UK at a fine resolution and due to the need to maintain equality of representation in political elections, have similar populations. CAS ward boundaries in particular have been the standard format for the release of ward level census information since 2003. They reflect electoral ward boundaries promulgated as at 31/12/2002 and contain 8850 separate wards for England and Wales [1, 2].

Census data used in the study was only available for England and Wales as the process for collating data in Scotland and the definition of geographic boundaries meant that equivalent datasets could not be produced. More information on census geography for Scotland for the 2001 census can be found at [3].

Datasets

This section outlines the details of the datasets used in our research. The source for each of the variables together with their correspondent code is described. When referring to census data, this corresponds to the 2001 UK census produced by the Office for National Statistics.

The original data and associated metadata for tables UV02, UV53, KS15, UV34, KS12 and household income can be found under the topics section of the UK neighbourhood statistics website [1].

Population (Table UV02)

The data on population was taken from UK census table UV02. Population data was taken from a data table on population density at the CAS ward level which provided separate statistics for total population, ward area and a result in population density figure. The total population figure was used for all regressions with socio-demographic variables used in the study.

Housing Stock (Table UV53)

Data on household dwelling numbers comes from census table UV53. The table provides information on the number of households, occupied or unoccupied, within each ward. Unoccupied household spaces are split into second residences/holiday accommodation, and vacant household spaces. A household space is the accommodation occupied by an individual household or, if unoccupied, available for an individual household. The population of this table is therefore all household spaces. The category used for regression was all household spaces and therefore included all spaces whether they were occupied or unoccupied.

Travel to Work (Table KS15)

Data on travel to work distances was taken from the UK census table KS15. The table shows both the length and the means of travel to work used for the longest part, by distance, of the usual journey to work. For the purposes of this table, public transport is defined as underground, metro, light rail or tram, train and bus, minibus or coach. The distance travelled to work is the distance in kilometres of a straight line between the residence postcode and workplace postcode. The distance is not calculated for people working mainly at or from home, people with no fixed workplace, people working on an offshore installation or people working outside the UK. The population of the table is all people aged 16 to 74 in employment.

Industry of Employment (Table UV34)

Data on the industry of employment of employees was taken from UK census table UV34. The table shows the usual resident population aged 16 to 74 in employment by the industry they work in. The industry in which a person works is determined by the response to the 2001 census question asking for a description of the business of the persons employer (or own business if self-employed). The responses were coded to a modified version of the UK Standard Industrial Classification of Economic Activities 1992 UK SIC (92).

In the 2001 census, industry of employment information was collected for usual residents. A usual resident was generally defined as someone who spent most of their time at a specific address. It includes: people who usually lived at that address but were temporarily away (on holiday, visiting friends or relatives, or temporarily in a hospital or similar establishment); people who worked away from home for part of the time; students, if it was their term-time address; a baby born before 30 April 2001 even if it was still in hospital; and people present on census day, even if temporarily, who had no other usual address. However, it did not include anyone present on census day who had another usual address or anyone who had been living or intended to live in a special establishment, such as a residential home, nursing home or hospital, for six months or more.

The industry of employment categories used in the study were the following: agriculture, hunting and forestry; manufacturing; construction; hotels and restaurants; financial intermediation; real estate, renting and business activities; public administration, defence and social security; education.

Occupational Groups (Table KS12a)

Data on occupational groups was taken from UK census table KS12. The information on this table comes from responses to questions asking for the full title of the main job and a description of the job. The population includes any person aged 16 to 74 who carried out paid work in the week before the census, whether self- employed or an employee, is described as employed or in employment. 'Paid work' includes casual or temporary work, even if only for one hour; being on a government- sponsored training scheme; being away from a job/business ill, on maternity leave, on holiday or temporarily laid off; or doing paid or unpaid work for their own or family business.

The following occupational groups were used in the regression analysis: managers and senior officials; professional occupations; associate professional and technical operations; skilled trades occupations; administrative and secretarial occupations; personal service occupations; sales and customer service occupations; process; plant and machine operatives; elementary occupations (examples of elementary occupations include farm workers, labourers, kitchen assistants and bar staff).

Household Income

The dataset on household income was taken from UK census experimental statistics for 2001/02 and is provided at a fine geographic resolution for the whole of England and Wales. The original data and associated metadata for household income can be found under the topics section of the UK neighbourhood statistics website [1].

The income dataset correspond to estimates that were produced using a model-based process which involves finding a relationship between survey data (data available on income) and other data drawn from administrative and census data sources. A model fitting process was used to select co-variates with a consistently strong relationship to the survey data. The strength of the relationship with these covariates was used to provide estimates on income for those wards where survey data on income is not available. More information on the provenance of the income data can be found on the appropriate page of the UK neighbourhood statistics census access site.

The survey data on income was taken from the Family Resources Survey (FRS) for the same year (2001/02)[1]. The total sample size for the 2001 survey was 42,000 addresses taken from

[1] The FRS is produced by the UK Department for Work and Pensions (DWP) to ensure a large sample sizes

across the UK.

In this study we used the average weekly household total income (unequivalised) estimations in Pounds Sterling. We converted the weekly average income of each ward into weekly sum of incomes by multiplying the average by the number of households. The number of households within each ward was taken from the census Household Composition table (UV65).

Patents

Patent data was provided by the UK Intellectual Property Office with postcode level reference that was subsequently aggregated to the CAS ward level. Data was provided for the years 2000 to 2011 inclusive to ensure sufficient quantity to avoid null values for individual wards. The total number of patents in the dataset that could be identified in E&W was 66,270. The values used for regression were simply the gross number of patents registered to a particular postcode whether it be business or home address.

More information on patent information from the UK IPO can be found at [8].

Land Use Statistics

All the measures related to land use were taken from the Generalised Land Use Database, 2001 (GLUD). Below we describe the geographical units and classifications.

The GLUD figures show the areas of different land types for census Output Areas (OAs), Lower Layer Super Output Areas (LSOAs), Middle Layer Super Output Areas (MSOAs), Local Authorities (LAs), and Government Office Regions (GORs) in England as at 1st November 2001. Output level data was aggregated to the ward level for comparative analysis with population.

For the GLUD, a classification has been developed which allocates all identifiable land features on the UK Ordnance Survey MasterMap national mapping product into nine simplified land categories and an additional 'unclassified' category. These are: (1) Domestic buildings (2) Non-domestic buildings (3) Roads (4) Paths (5) Rail (6) Gardens (domestic) (7) Greenspace (8) Water (9) Other land uses (largely hardstanding) and (10) Unclassified.

Scaling results for cities defined at $\rho_c = 14$prs/ha

This section gives a summary of the results presented for cities and metropolitan areas in the main text. In Fig. 3 we show that for the choice of $\rho_c = 14$prs/ha, the system of cities obtained is in very good agreement with the identified urbanised space observed through satellite images. The cumulative distribution for population is given by Fig. 3(b), and it clearly obeys

when collating information on household expenditure. Information on the FRS for 2001 can be found on the research section of the dWP.gov website [6], the methodology section (section 8) of the FRS summary report for that year at [7] and the associated technical report available through the ONS [2]

Zipf's law. Fig. S1 shows the same figure for metropolitan areas, i.e. for $\rho_c = 14$prs/ha and 30% commuters. We observe that Zipf's law still holds, and both distribution are very similar. Table S1 presents the results of the analysis carried out for 30 different variables, for cities and metropolitan areas defined at $\rho_c = 14$prs/ha. These results give the values and confidence intervals for the indicators presented in the main text in Fig. 5.

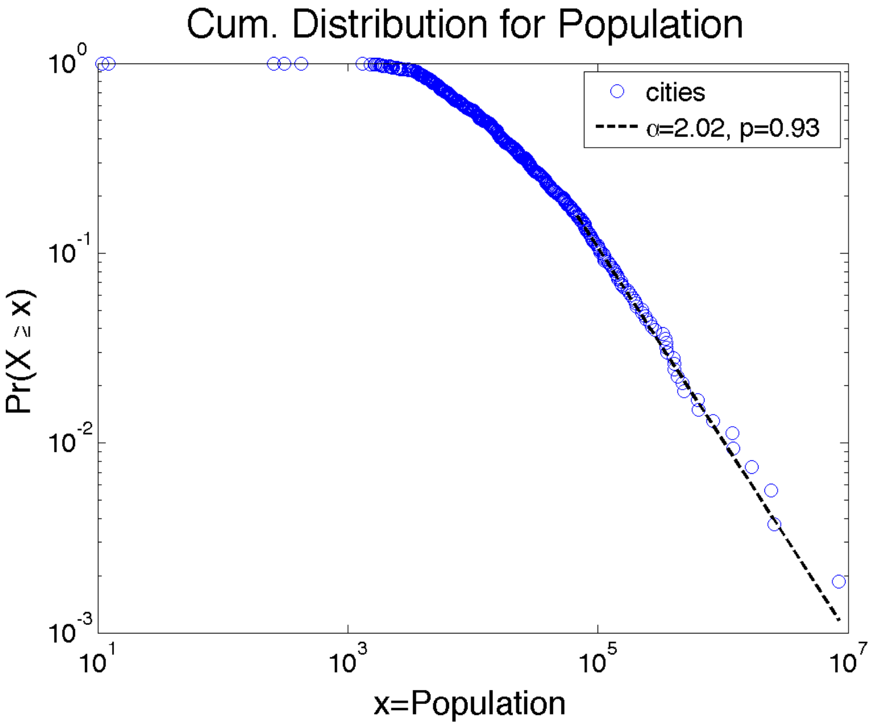

Figure S1: Zipf distribution of metropolitan areas for $\rho_c = 14$prs/ha and 30% commuters.

Sensitivity of the scaling exponent to city definitions

In this section we extend the results of the sensitivity analysis for the scaling exponent to a dozen urban indicators. We show the results for no population size cutoff, and for a cutoff of 10^4 individuals. The behaviour of the exponent is displayed through heatmaps covering the full parameter space. On the horizontal axis there are 40 different population density thresholds ($\rho = 1, 2, ..., 40$prs/ha) and on the vertical axis there are 21 different percentages for commuting thresholds ($\tau = 0, 5, ..100\%$ individuals).

Let us first look at exponents of variables that should lie in the sublinear regime Fig. S2. We observe nevertheless, that for two such important infrastructure variables, area of roads and area of paths, the exponent is not at all sublinear. From this selection, only distance to work truly

Table S1: Urban indicators for cities defined at a density of 14prs/ha with and without commuters

Urban indicators for $\rho = 14$prs/ha	no commuting			30% commuters			
UK census 2001	β	95% CI	R^2	β	95% CI	R^2	n. cities
Distance to work (km)	0.78	[0.76,0.81]	0.86	0.82	[0.80,0.85]	0.88	535
Empl. agriculture, hunting, forestry	0.86	[0.83,0.89]	0.84	0.92	[0.89,0.95]	0.87	535
Area of greenspace (1000m^2)*	0.95	[0.92,0.99]	0.85	1.23	[1.19,1.28]	0.86	477
Area of domestic gardens (1000m^2)*	0.97	[0.95,0.99]	0.96	1	[0.99,1.02]	0.96	477
Area of dom. buildings (1000m^2)*	0.97	[0.96,0.98]	0.99	0.99	[0.98,1.00]	0.99	477
Empl. construction	0.98	[0.97,1.00]	0.97	0.99	[0.98, 1.00]	0.98	535
Empl. manufacturing	1	[0.97,1.02]	0.93	1	[0.97,1.02]	0.94	535
Area of road (1000m^2)*	1	[0.99,1.01]	0.99	1.04	[1.03,1.06]	0.98	477
Empl. process; plant & machine op.	1.01	[0.98,1.03]	0.93	1	[0.98,1.03]	0.94	535
Empl. elementary occupations	1.01	[0.99,1.02]	0.98	1	[0.99,1.01]	0.98	535
Area of rail (1000m^2)*†	1.03	[0.88,1.18]	0.67	1.05	[0.90,1.20]	0.67	97
Area of path (1000m^2)*	1.06	[1.02,1.09]	0.88	1.06	[1.03,1.09]	0.89	477
All household spaces	0.99	[0.98,0.99]	1	0.99	[0.98,0.99]	1	535
Empl. personal service occupations	0.99	[0.98,1.00]	0.99	0.99	[0.98,1.00]	0.99	535
All people aged 16-74 in employment	1	[0.99,1.01]	0.99	1	[1.00, 1.01]	0.99	535
Total number of patents 2000-2011†	0.98	[0.84,1.13]	0.67	0.99	[0.86,1.12]	0.71	100
Number of train stations†	0.98	[0.86,1.11]	0.73	1.04	[0.91,1.16]	0.75	100
Empl. hotels and restaurants	0.98	[0.97,1.00]	0.96	0.98	[0.97,1.00]	0.96	535
Empl. skilled trades occupations	0.99	[0.97,1.00]	0.98	0.99	[0.98,1.00]	0.98	535
Empl. managers and senior officials	0.99	[0.97,1.02]	0.94	1.01	[0.99,1.03]	0.95	535
Net income (weekly) (before housing)	0.99	[0.98,1.00]	0.98	1	[0.99,1.01]	0.99	535
Net income (weekly) (after housing)	0.99	[0.98,1.01]	0.98	1	[0.99,1.01]	0.99	535
Total income (weekly)	1	[0.99,1.02]	0.98	1.01	[0.99,1.02]	0.98	535
Net income (weekly)	1	[0.99,1.02]	0.98	1.01	[1.00,1.02]	0.98	535
Empl. ass. prof. and technical occ.	1.01	[1.00,1.03]	0.96	1.02	[1.00,1.03]	0.96	535
Empl. professional occupations	1.03	[1.00,1.05]	0.9	1.03	[1.01,1.06]	0.91	535
Empl. real estate, business activities	1.03	[1.01,1.06]	0.93	1.03	[1.01,1.05]	0.93	535
Empl. sales, customer service occ.	1.03	[1.02,1.04]	0.99	1.02	[1.01,1.03]	0.99	535
Empl. financial intermediation	1.08	[1.04,1.11]	0.86	1.08	[1.04,1.11]	0.88	535
Area non-dom. buildings (1000m^2)*	1.11	[1.08,1.15]	0.88	1.13	[1.09,1.17]	0.89	477

* Welsh cities are excluded, since data for Wales on infrastructure is not available

† Only considered cities with a minimum population size of 5.10^4 individuals in order to avoid null values

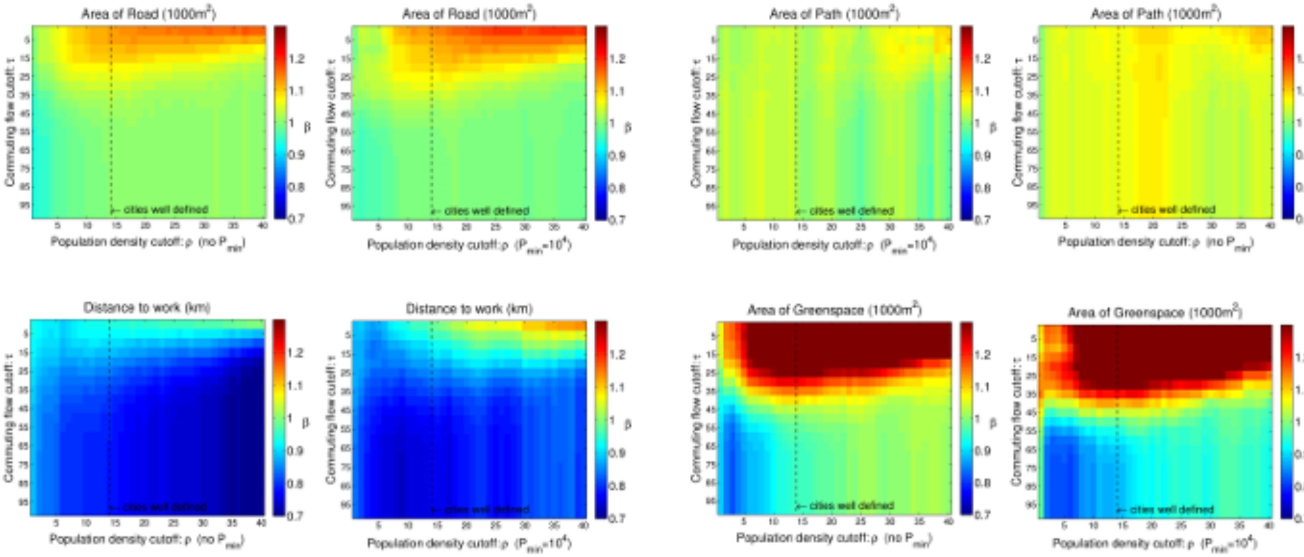

Figure S2: Heatmaps for variables whose exponent is expected to behave as $\beta < 1$, such as infrastructure variables. For each variable we have 2 maps: one with no minimum population size, and another for a minimum population size of 10^4 individuals.

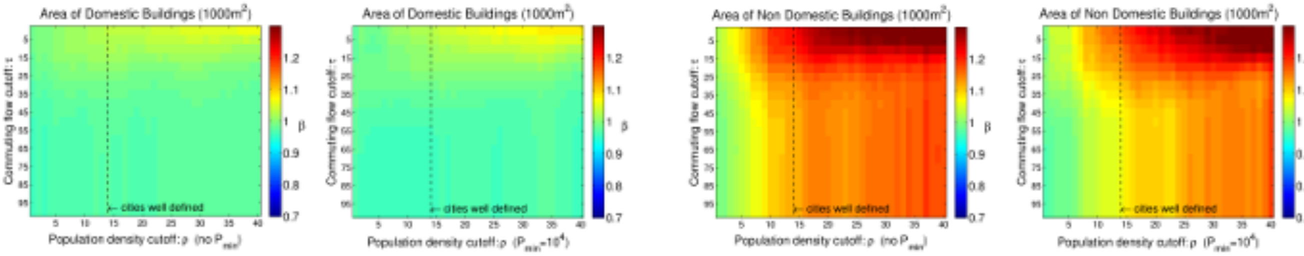

Figure S3: Heatmaps for area of domestic and non-domestic buildings.

belongs to the sublinear regime. On the other hand, area of green space clearly demonstrates the sensitivity of the exponent to city boundaries when nonlinear dependencies are present.

Further cases of interest within the infrastructure variables are the area of domestic and of non-domestic buildings, Fig. S3. The latter can be seen as reflecting the economic activity within cities, and once again we see how the exponent varies depending on the boundaries and distribution of cities considered.

Let us now turn our attention to variables corresponding to employment categories. Fig. S4 shows the heatmaps for categories expected to lie within the sublinear regime, since these correspond to employment requiring only basic skills. On the other hand, Fig. S5 depicts the categories where skilled individuals are needed.

For basic employment categories, only employment in agriculture belongs to the sublinear regime. The heatmaps show that the value of the exponent is sensitive to commuting flows.

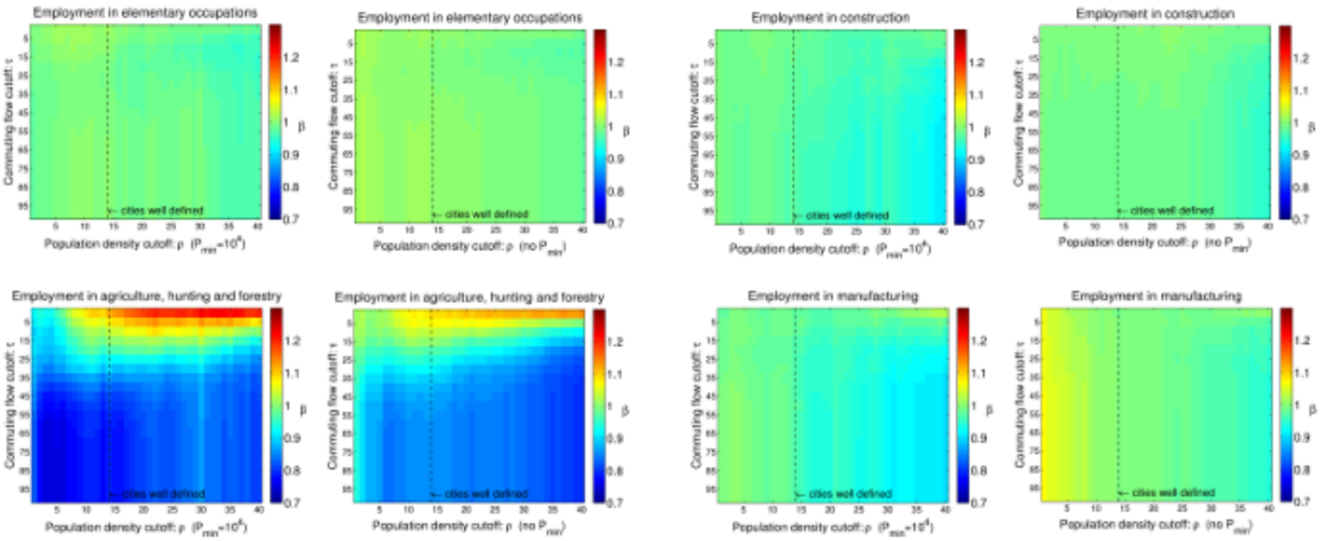

Figure S4: Heatmaps for basic employment categories.

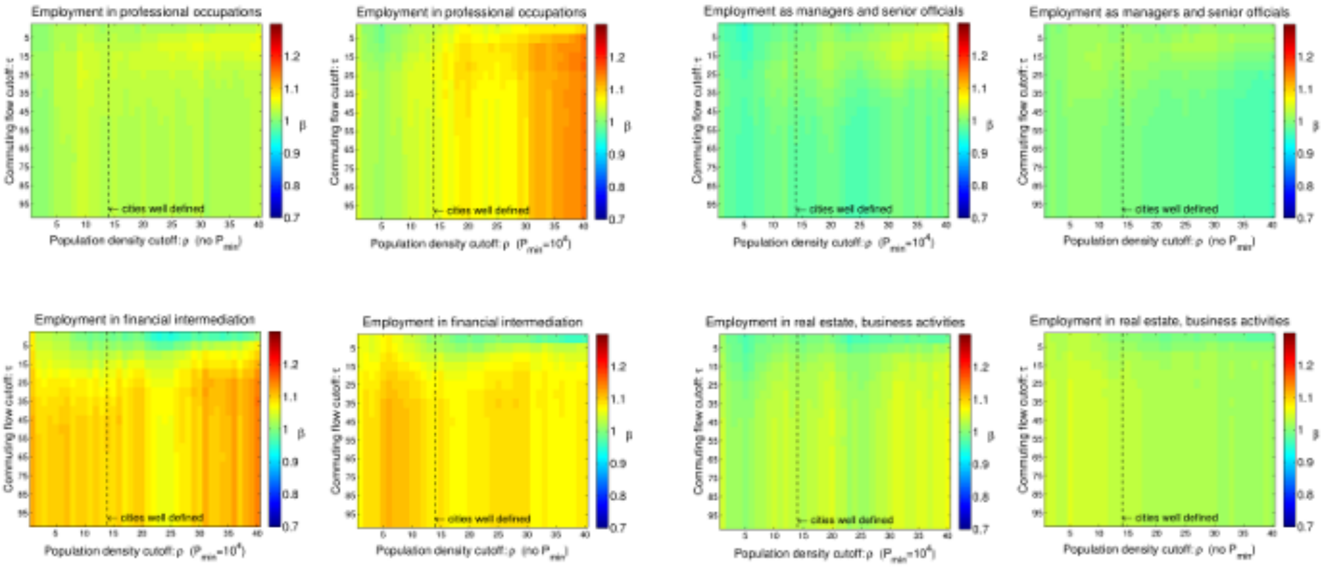

Figure S5: Heatmaps for skilled employment categories.

For employment in professional occupations and financial intermediation, mild superlinearity is observed. Nevertheless, the exponent seems to belong to the linear regime for employment as managers and in real estate. It is important to note that for employment in professional occupations, a sensitivity for city boundaries is present for a population cutoff of 10^4 but none is apparent when no cutoff is applied.

Dragon-Kings

We argued in the text that size alone does not provide enough information to define the state of a city according to eq. (1). The scaling effects observed in certain countries, might pertain to a very specific system of cities not found in E&W. In this particular case, London behaves as an outlier. This is an extremely important city that cannot be removed from the statistics of cities, as one might proceed when encountering outliers in a distribution classified as errors or biases in the data. In this case, we are observing an outlier whose dynamics are different to the rest of the cities in the distribution. These particular sets of large events are identified by Sornette and collaborators (see references in main text) as *dragon-kings*, given that they are extremely important (kings), and different in nature to the rest of the events in the distribution (dragons). These might be the outcome of some sort of amplification mechanism from positive feedbacks.

There is no unique methodology to identify these sorts of events. Here we visually show that London is an outlier by employing the transformation in [4]. Given that a power law is a special case of a stretched exponential distribution, we transform the size of cities s as follows:

$$s \rightarrow \exp - \left(\frac{s}{s_0} \right)^c \tag{S1}$$

The usual rank size distribution of city sizes is shown in Fig. S6 (left), and appears to be more or less Zipf. The distribution using the transformation in eq. (S1) is given in Fig. S6 (right) which clearly shows London as a red point well outside the fitting line.

Further statistical tests showing that London is an outlier can be found in [5]. There they employ two different tests, the DK and U-test, which consist in showing that London, as the larger event, is not generated by the same power law distribution as the rest of the cities.

Scaling results for Larger Urban Zones (E&W)

In order to provide a wider perspective of the behaviour of urban indicators beyond our own definition of cities, we present the results for the definition of cities in E&W in terms of Larger Urban Zones (LUZ). LUZ were introduced by Eurostat in order to provide a consistent definition of cities across Europe. A map of the LUZ cities and a plot of thier size distribution are presented in Fig. S7.

The analysis is undertaken by classifying each indicator in terms of the expected three regimes: sublinear, linear and superlinear. The results are shown in Fig. S8, and the details

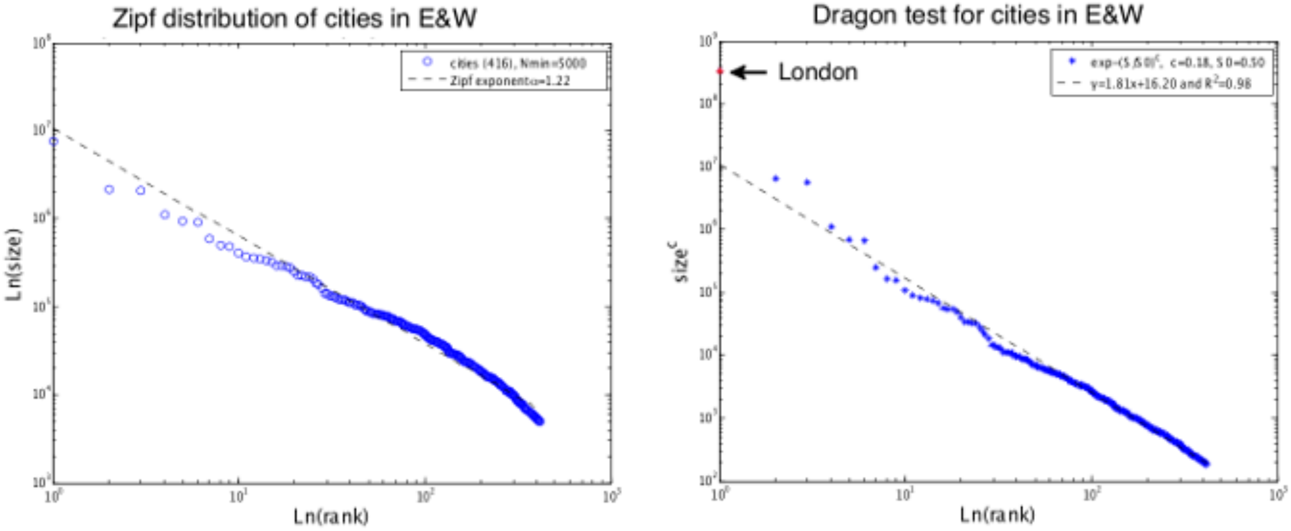

Figure S6: Cities in E&W are defined at $\rho_c = 14$prs/ha. Left: Zipf's law distribution of cities; right: transformed distribution showing London as a dragon-king.

of the variables and their statistical properties can be found in Table S2. We observe that once again the expected scaling behaviour is not corroborated for the urban indicators lying in the superlinear regime.

References

[1] http://www.neighbourhood.statistics.gov.uk/dissemination/.

[2] http://www.ons.gov.uk/.

[3] http://www.gro-scotland.gov.uk/statistics/.

[4] Sornette D. Dragon-Kings, Black Swans and the Prediction of Crises. Int J Terraspace Sci Eng. 2009;2:1–18.

[5] Pisarenko VF, Sornette D. Robust Statistical Tests of Dragon-Kings beyond Power Law Distributions. Eur Phys J-Spec Top. 2012;205:95–115. Available from: http://arxiv.org/abs/1104.5156.

[6] http://research.dwp.gov.uk/asd/frs/2001_02/index.php?page=intro.

[7] http://research.dwp.gov.uk/asd/frs/2001_02/pdfonly/frs_2001_02_report.pdf.

[8] http://www.ipo.gov.uk.

Table S2: Urban indicators for LUZ in E&W

Urban indicators for LUZ, UK census 2001	β	95% CI	R^2	n. cities
Employed in agriculture, hunting and forestry	0.59	[0.35,0.83]	0.58	21
Distance to work (km)	0.70	[0.51,0.88]	0.76	21
Area of road (1000m^2)*	0.82	[0.71,0.93]	0.93	19
Area of domestic gardens (1000m^2)*	0.84	[0.75,0.93]	0.96	19
Employed in manufacturing	0.90	[0.79,1.02]	0.93	21
Number of bus stops	0.91	[0.82,1.01]	0.96	21
Area of path (1000m^2)*	0.94	[0.84,1.05]	0.96	19
Employed: process; plant and machine op.	0.95	[0.84,1.05]	0.95	21
Employed in elementary occupations	0.95	[0.91,0.99]	0.99	21
Area of rail (1000m^2)*	0.95	[0.76,1.15]	0.86	19
All people aged 16-74 in employment	0.99	[0.96,1.02]	1.00	21
All household spaces	1.00	[0.99,1.01]	1.00	21
Consumption of domestic electricity	1.06	[0.95,1.16]	0.96	21
Employment in skilled trades occupations	0.92	[0.86,0.98]	0.98	21
Total number of patents (2000-2011)	0.95	[0.64,1.26]	0.68	21
Employed in hotels and restaurants	0.97	[0.94,1.01]	0.99	21
Employed in professional occupations	1.01	[0.89,1.12]	0.94	21
Employed as managers and senior officials	1.01	[0.92,1.09]	0.97	21
Employment in associate prof and technical occ.	1.01	[0.94,1.08]	0.98	21
Total income (weekly)	1.03	[0.96,1.10]	0.98	21
Employed in real estate, business activities	1.06	[0.93,1.20]	0.94	21
Number of train stations	1.12	[0.86,1.38]	0.81	21
Employed in financial intermediation	1.25	[1.13,1.36]	0.96	21

* The two Welsh cities in LUZ are excluded, since data for Wales on infrastructure is not available

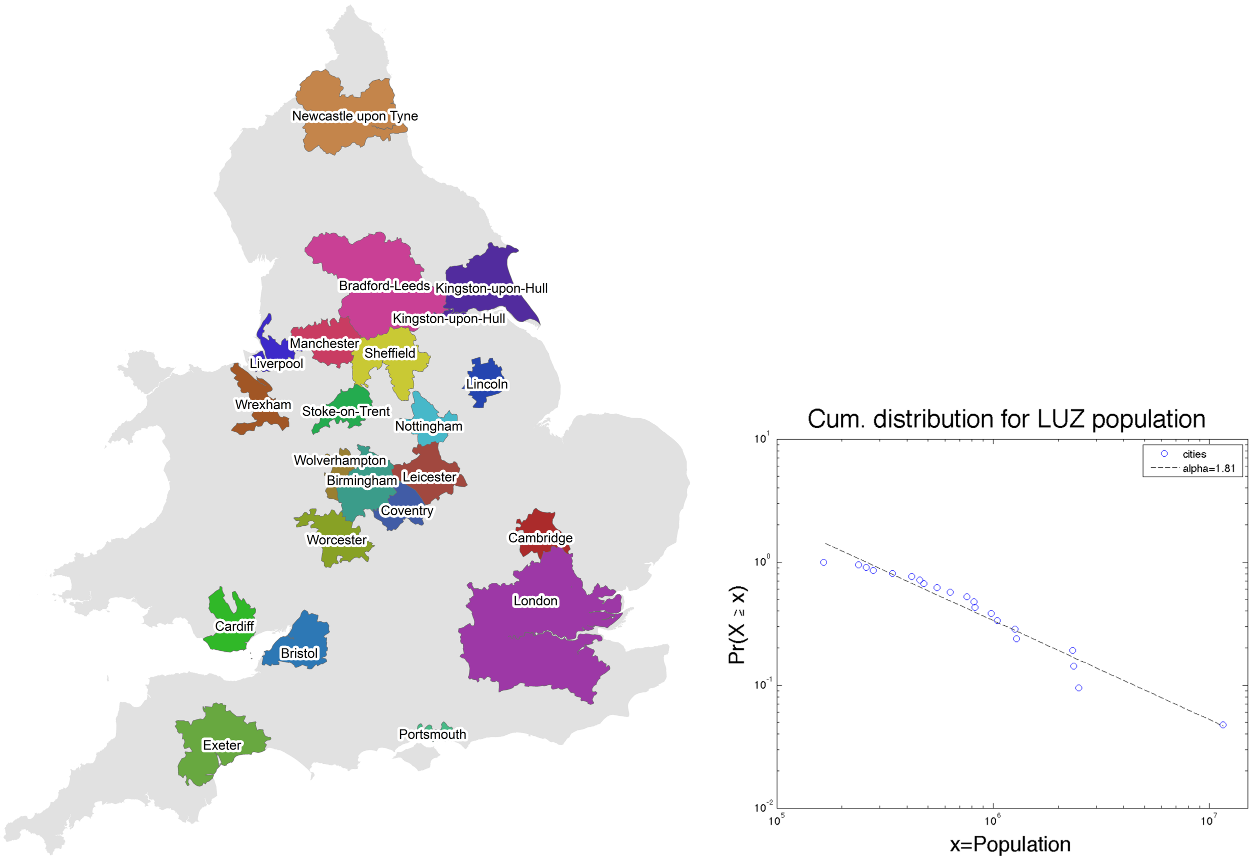

Figure S7: Map (left) and size distribution (right) of LUZ for E&W

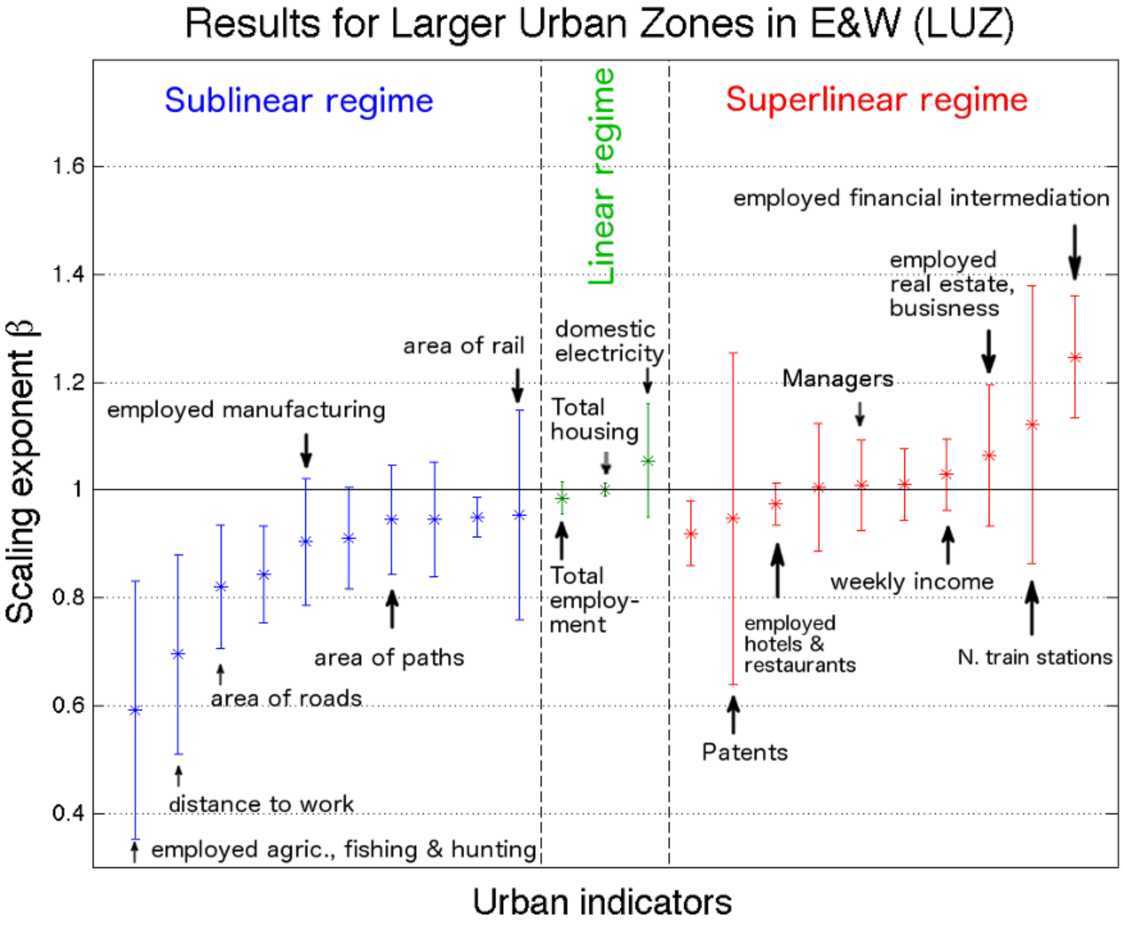

Figure S8: Exponents with 95% CI for different urban indicators coloured-coded according to their expected regime.