Supplementary information to "Quantifying ethnic segregation in cities through random walks"

Sandro Sousa[1] and Vincenzo Nicosia[1]

[1] *School of Mathematical Sciences, Queen Mary University of London, London E1 4NS, United Kingdom*

CORRELATIONS AT MULTIPLE SPATIAL SCALES

In addition to the class cover time analysis, we look at correlations among the nodes of G at different spatial scales. We adopted the multifractal detrended fluctuation analysis (DFA), a non-linear method which allows to detect the presence of long-term correlations of a time series [1]. With a a slightly different set-up, random walks are defined on the graph G representing the UK cities at three spatial scales (geographic delineations), the Wards level, Lower Layer Super Output Areas (LSOA) and Output Areas (OA) respectively. The choice for φ_i is the Shannon entropy of the population distribution living in an area, namely the node population entropy $\varphi_i \equiv x_i = S_i$. Then, for every node i in G, each ethnic group γ is obtained from the UK Census data ($C = 250$ different classes) and it is defined as:

$$p_i = p_i^\gamma = \frac{x_i}{\sum_\gamma x_i^\gamma}, \tag{1}$$

and the node population entropy is given by:

$$x_i \equiv S_i = \sum_{\gamma=1}^{C} p_i^\gamma \log\left[p_i^\gamma\right]. \tag{2}$$

This walk produces the time series $y_i(t) = (S_{i1}, S_{i2}, S_{i3}, ...)$ of length T containing the node population entropy. The length T is fixed and defined according to the number of nodes in G and the spatial scale. Given $Y_i(t)$, the cumulative sum, or profile, is obtained by subtracting the mean value $\langle Y\rangle$ of the time series:

$$X_t \equiv \sum_{i=1}^{t}(Y_i - \langle Y\rangle), \quad t = 1, ..., T \tag{3}$$

Next, the profile X_t is divided in non-overlapping segments of equal length ε and the local trend is calculated by a least-square linear fit of each segment. Let K_t be the resulting piece of this process, then, the root-mean-square fluctuation of this detrended time series is calculated by:

$$F(\varepsilon) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(X_t - K_t)^2} \tag{4}$$

The local variance $\sigma^2(\ell, \varepsilon)$ is obtained and the structure function $F(\varepsilon)$ is evaluated by averaging the $\sigma^2(\ell, \varepsilon)$ over all time windows whose length equals ε and $F(\varepsilon)$ is plotted as a function of ε.

If the probability to find the edge (i, j) connecting node i to node j does not depend on the values φ_i and φ_j, then the fluctuations of the corresponding time series $Y_i(t)$ obtained from a random walk will be indistinguishable from an uncorrelated Gaussian noise $F(\varepsilon) \approx \varepsilon^{1/2}$. On the other hand, $F(\varepsilon) \approx \varepsilon^\alpha$ with $\alpha \neq 1/2$ signals the presence of φ-correlations. For all UK cities in Fig. 12, φ-correlations can be observed in the regime below the transition point where the scaling exponent α shows the magnitude of such correlation. Above the transition point, all cities show fluctuations on $Y_i(t)$ indistinguishable from uncorrelated Gaussian noise. These two scaling regimes indicate that the network looks different in respect to the node property when observed at local and global scale.

At small values of ε the walker is exploring the network for relative short time intervals (local scale) and it observes correlated fluctuations in the node population entropy (green fit line in Fig. 12), while at large values (blue fit line) the network appears uncorrelated ($\alpha \approx 1/2$). The transition point marked by the vertical dashed line separating the two regimes corresponds to the typical walk length ε above which local heterogeneities and correlations in the values of $Y_i(t)$ are less significant. Therefore, for any length up to the transition point, the walker observes correlations on the fluctuations of φ, suggesting that there is no need to move too far on G to find the spatial patterns governing the population distribution.

Notably, in the case of London, the coverage of the neighbourhoods visited by W corresponds to an average area of 778, 779 and 796 Km^2 for Wards, LSOA and OA respectively. The areas correspond to the walk length at the transition point where London has a total area of 1572 Km^2. The distances are obtained by splinting $Y_i(t)$ on non overlapping segments, obtaining the total area covered at each segment and computing the average over all segments. Detailed values for all UK cities can be observed in Table I. The mean covered area is relatively constant and indicates that the walker can extract the information about the correlations among the nodes of G no matter which spatial scale is used to delineate the territory.

TABLE I. Table showing results from the DFA for the main metropolitan areas in the UK at the spatial scales of Wards, LSOA and OA respectively. For all cities, the walks were simulated with lengths 1e+6, 50e+6 and 300e+6 for Wards, LSOA and OA respectively, except London where walks were defined with lengths 2e+6, 100e+6 and 900e+6 respectively.

Met. area	Scale	Walk length	Slope h	Slope t	Cut-off x	Cut-off y	Mean area	Std	Mean length	Std
Bristol	Wards	4e+6	1.05	0.54	1.92	0.44	459.91	162.99	42.22	12.39
Bristol	LSOA	50e+6	1.13	0.54	2.67	1.21	465.05	183.05	41.5	12.3
Bristol	OA	300e+6	1.11	0.52	3.61	2.09	527.35	168.57	46.44	11.88
Cardiff	Wards	4e+6	1.01	0.5	2.67	0.98	1469.95	316.39	62.79	9.34
Cardiff	LSOA	50e+6	1.09	0.51	3.42	1.84	1952.08	385.57	69.17	7.89
Cardiff	OA	300e+6	1.09	0.52	3.99	2.38	1598.51	366.2	65.51	8.93
Liverpool	Wards	2e+6	1.0	0.55	2.11	0.46	304.57	77.61	31.35	6.52
Liverpool	LSOA	50e+6	1.1	0.52	2.86	1.27	224.37	59.37	28.49	6.37
Liverpool	OA	300e+6	1.08	0.51	3.8	2.12	318.34	72.56	33.72	6.46
London	Wards	6e+6	1.03	0.52	3.05	1.57	778.55	131.84	47.13	6.78
London	LSOA	100e+6	1.07	0.53	3.99	2.49	779.48	115.31	48.41	6.47
London	OA	900e+6	1.06	0.52	4.74	3.19	796.28	111.35	49.49	6.16
Manchester	Wards	4e+6	0.99	0.55	2.29	0.84	421.75	89.02	37.08	7.04
Manchester	LSOA	50e+6	1.08	0.55	3.24	1.77	427.8	86.87	37.85	6.83
Manchester	OA	300e+6	1.09	0.55	3.99	2.49	443.46	85.54	39.37	6.91
Northeast	Wards	2e+6	0.96	0.53	2.29	0.58	2270.74	1508.25	74.79	25.81
Northeast	LSOA	50e+6	1.07	0.52	2.86	1.2	1636.09	1522.99	63.96	31.42
Northeast	OA	300e+6	1.05	0.53	3.8	2.04	2254.1	1591.52	78.79	31.4
Sheffield	Wards	2e+6	0.86	0.5	2.67	0.98	1396.21	123.98	61.71	3.03
Sheffield	LSOA	50e+6	1.02	0.53	3.24	1.66	939.55	190.49	53.34	8.52
Sheffield	OA	300e+6	1.05	0.52	3.99	2.36	914.46	186.18	53.75	8.49
WMidlands	Wards	4e+6	0.99	0.53	1.92	0.56	200.85	66.02	25.41	6.65
WMidlands	LSOA	50e+6	1.11	0.53	3.05	1.68	197.82	52.11	26.89	6.51
WMidlands	OA	300e+6	1.1	0.53	3.99	2.53	282.34	63.33	32.5	7.31
WYorkshire	Wards	4e+6	0.85	0.51	2.29	0.71	1088.52	211.34	52.66	6.9
WYorkshire	LSOA	50e+6	1.07	0.53	2.86	1.4	428.86	137.81	38.28	8.39
WYorkshire	OA	300e+6	1.08	0.53	3.61	2.1	442.01	132.49	39.99	8.34

[1] Kantelhardt, J.W., Zschiegner, S.A., Konscienly-Bunde, E., Havlin, S., Bunde, A. and Stanley, H.E. (2002) Multifractal detrended fluctuation analysis of nonstationary time series. Physica A: Statistical Mechanics and Its Applications, 316, 87-114. doi:10.1016/S0378-4371(02)01383-3

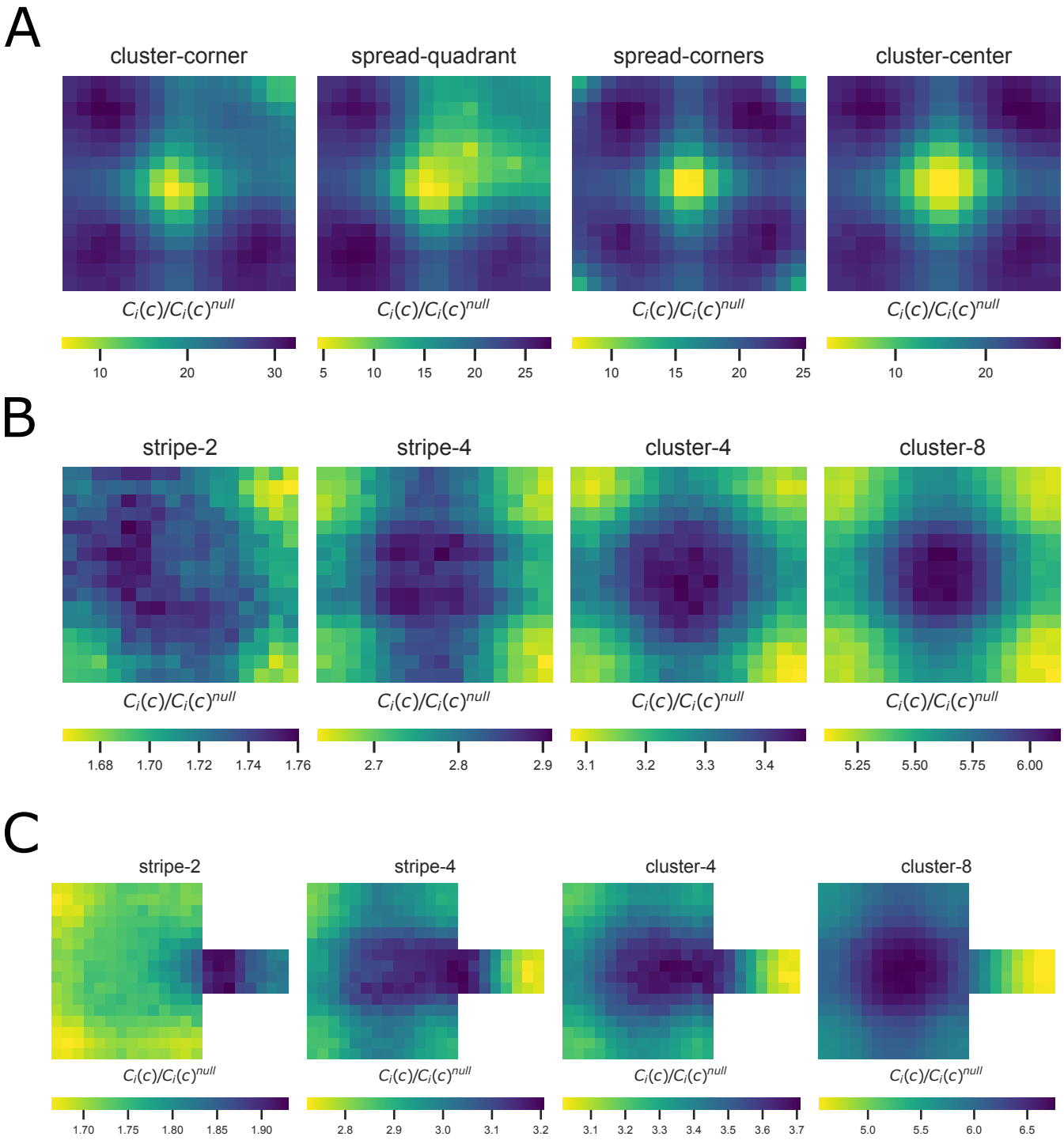

FIG. 1. Heatmaps reporting the normalised coverage times $C_i(c)/C_i(c)^{null}$ for $c = 0.7$ of the synthetic colourings of lattices in Fig.3 of the main text. (A) The population is divided in 5 classes displaced over 4 different set-ups. (B) and (C) 32 classes are placed randomly with 4 different setups on on a lattice with a lateral appendix.

TABLE II. Properties of the metropolitan areas in the US and UK considered in this work. Nodes and links correspond to the spatially-embedded graph constructed from the neighbourhoods adjacency at the corresponding spatial scale of each country. Population data according to Census 2011 for the UK AND Census 2010 for US.

Met. area	Scale	Population	Classes	Nodes	Links	$\Delta\varrho$	$\Delta\sigma$	$\Delta\mu$
Atlanta	Census	5,618,431	60	1019	3172	311.23	15.68	1036.11
Boston	Census	7,558,009	63	1684	5130	667.81	23.35	4644.68
Chicago	Census	9,686,021	60	2273	7504	640.18	11.88	1317.09
Dallas	Census	6,726,779	60	1394	4466	479.62	11.03	1485.44
Houston	Census	6,045,555	62	1096	3603	441.93	13.18	1323.22
Los Angeles	Census	17,872,910	64	3923	12704	639.40	23.52	833.26
New York	Census	22,085,649	63	5277	16669	743.40	12.43	1666.32
Philadelphia	Census	6,533,683	60	1602	4940	452.69	11.23	840.99
San Francisco	Census	7,468,390	63	1651	5256	351.52	11.11	552.33
Washington	Census	8,572,971	63	2082	6600	613.15	10.57	1054.84
Bristol	Wards	1,069,583	227	143	402	228.49	16.84	646.96
Cardiff	Wards	1,480,251	222	287	801	612.37	17.15	2590.53
Liverpool	Wards	1,506,935	225	132	350	244.53	11.21	775.83
London	Wards	8,173,941	250	632	1859	49.03	21.08	268.49
Manchester	Wards	2,682,528	246	215	603	184.06	13.65	533.56
North East	Wards	1,934,095	225	241	669	308.14	20.02	985.77
Sheffield	Wards	1,343,601	226	91	245	177.64	13.62	478.73
West Midlands	Wards	2,736,460	249	163	436	157.02	14.42	357.70
West Yorkshire	Wards	2,226,058	239	124	345	136.36	13.84	194.42

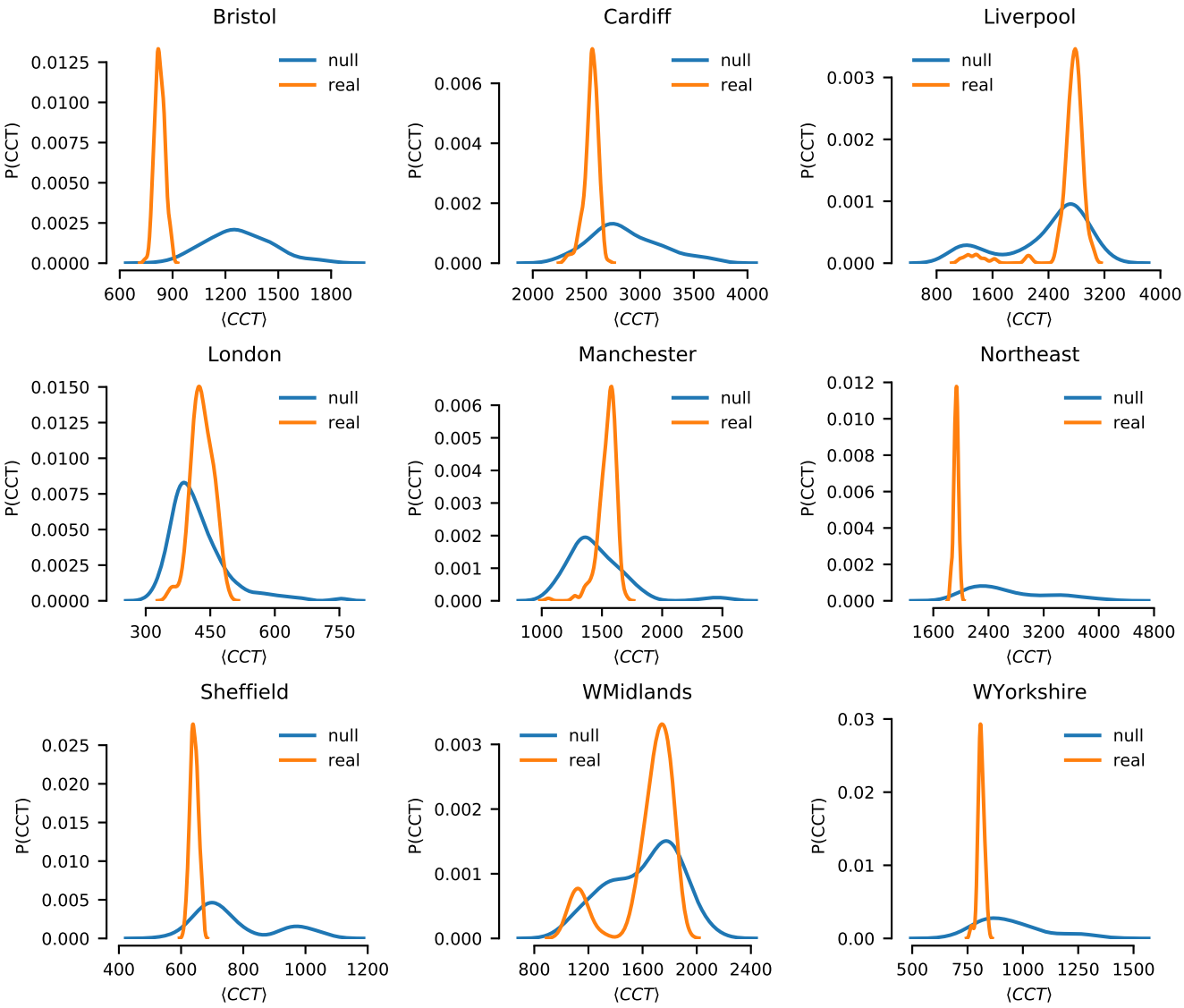

FIG. 2. Distribution of the Class Coverage Time across the wards of metropolitan areas in the UK, where larger values of CCT correspond to more segregated areas. The estimates for CCT are obtained using 1.000 distinct trajectories from each node. The observed distributions are qualitatively and quantitatively distinct from the corresponding distribution in a null-model where the profiles of classes is reshuffled uniformly at random (100 different realisations). The population is divided into $\Gamma = 250$ different classes obtained from the UK Census.

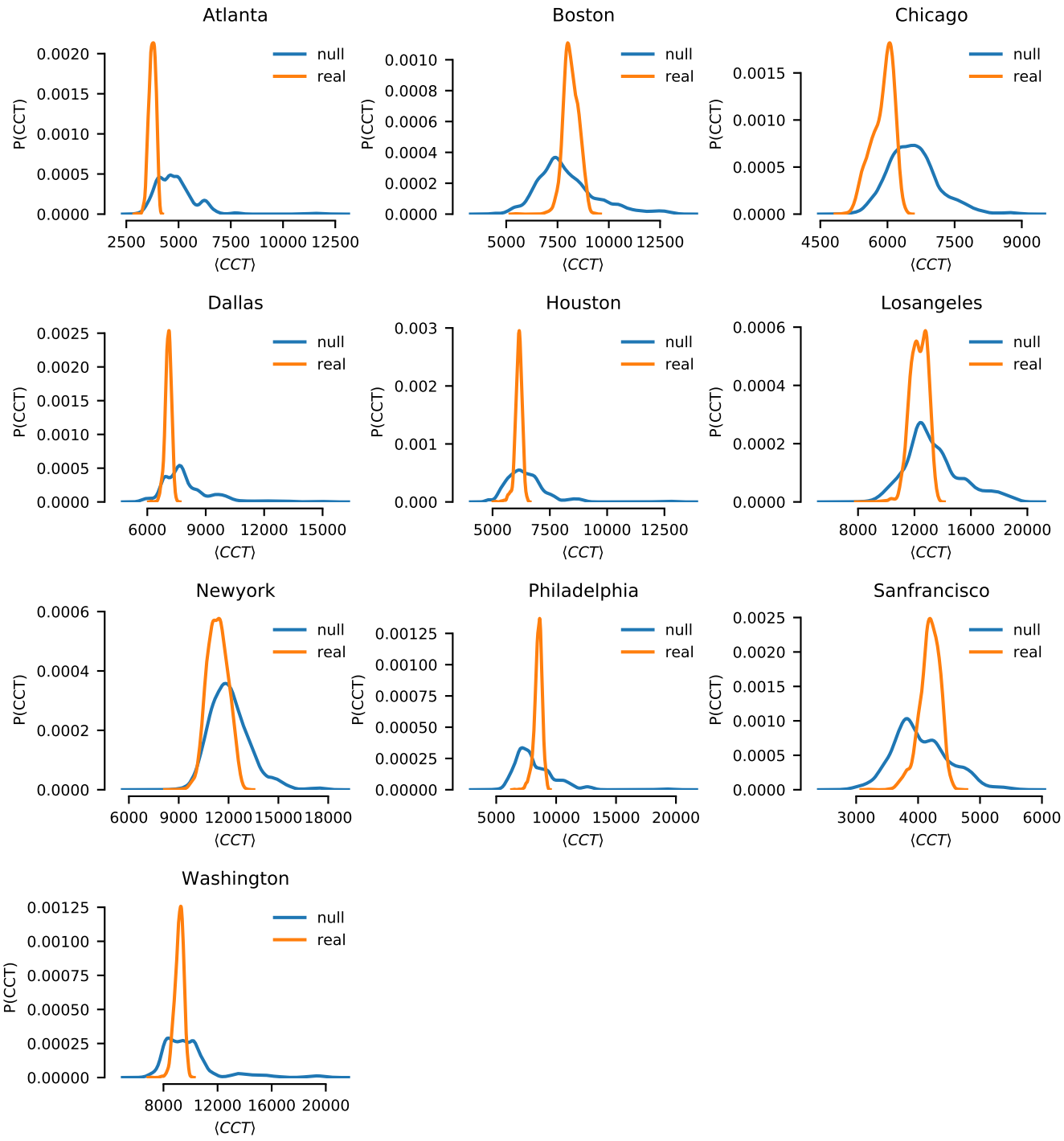

FIG. 3. Distribution of the Class Coverage Time across the wards of metropolitan areas in the US, where larger values of CCT correspond to more segregated areas. The estimates for CCT are obtained using 1.000 distinct trajectories from each node. As in the UK cities, the observed distributions are qualitatively and quantitatively distinct from the corresponding distribution in a null-model where the profiles of classes is reshuffled uniformly at random (100 different realisations). The population is divided into $\Gamma = 64$ different classes obtained from the American Census Bureau.

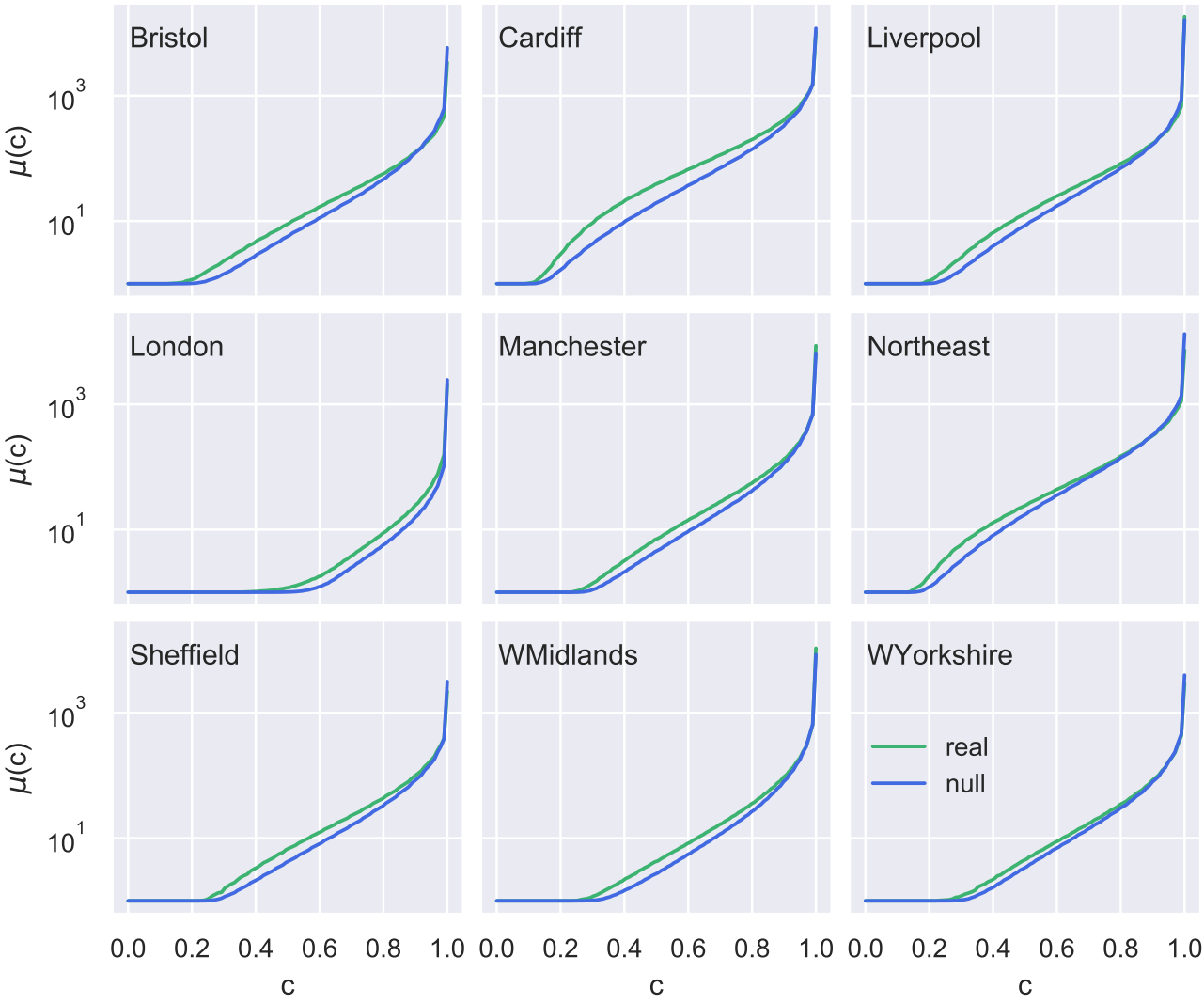

FIG. 4. Mean time $\mu(c)$ as a function of fractions c of classes in the real system and the corresponding null model for the metropolitan areas in the UK.

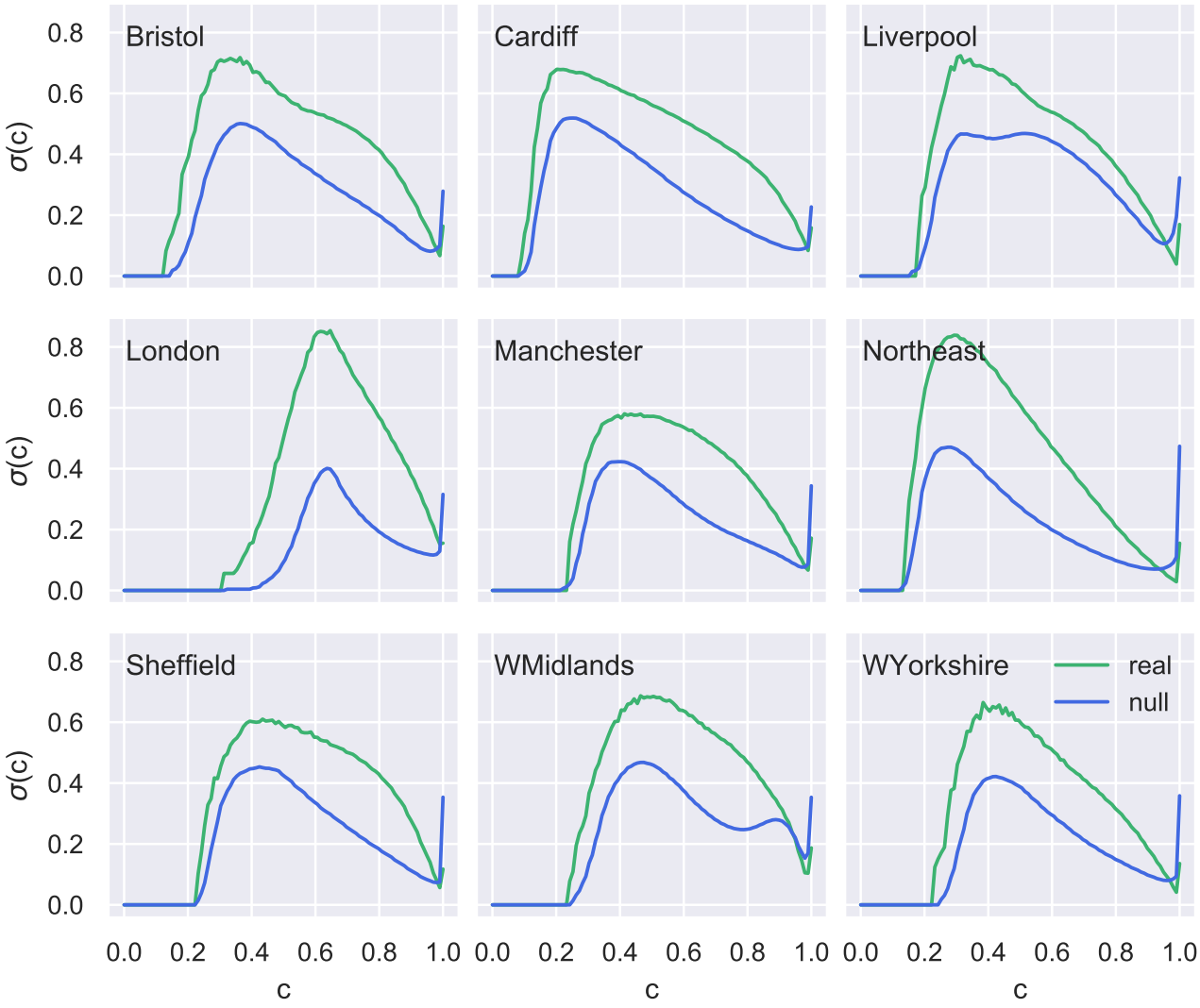

FIG. 5. Coefficient of variance $\sigma(c)$ as a function of fractions c of classes in the real system and the corresponding null model for the metropolitan areas in the UK. The values of $\sigma(c)$ are normalised by the respective mean $\mu(c)$. Values for $c = 1$ are kept for demonstration of the spurious effects. The flat line for small values of c is due to the number of classes observed at the initial node i being larger than the fraction c considered.

FIG. 6. Spatial diversity $\varrho(c)$ as a function of fractions c of classes in the real system and the corresponding null model for the metropolitan areas in the UK.

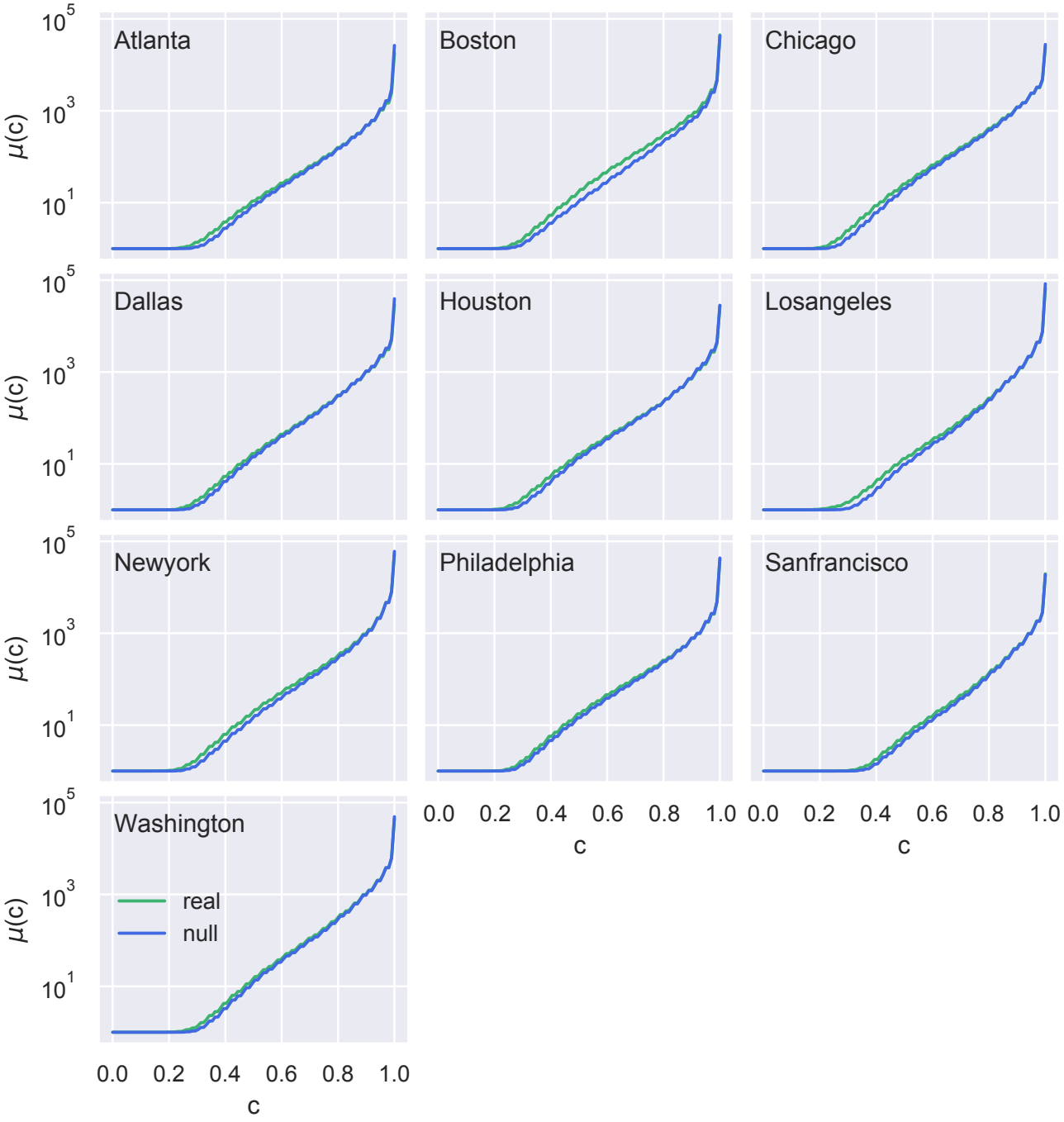

FIG. 7. Mean time $\mu(c)$ as a function of fractions c of classes in the real system and the corresponding null model for the metropolitan areas in the US.

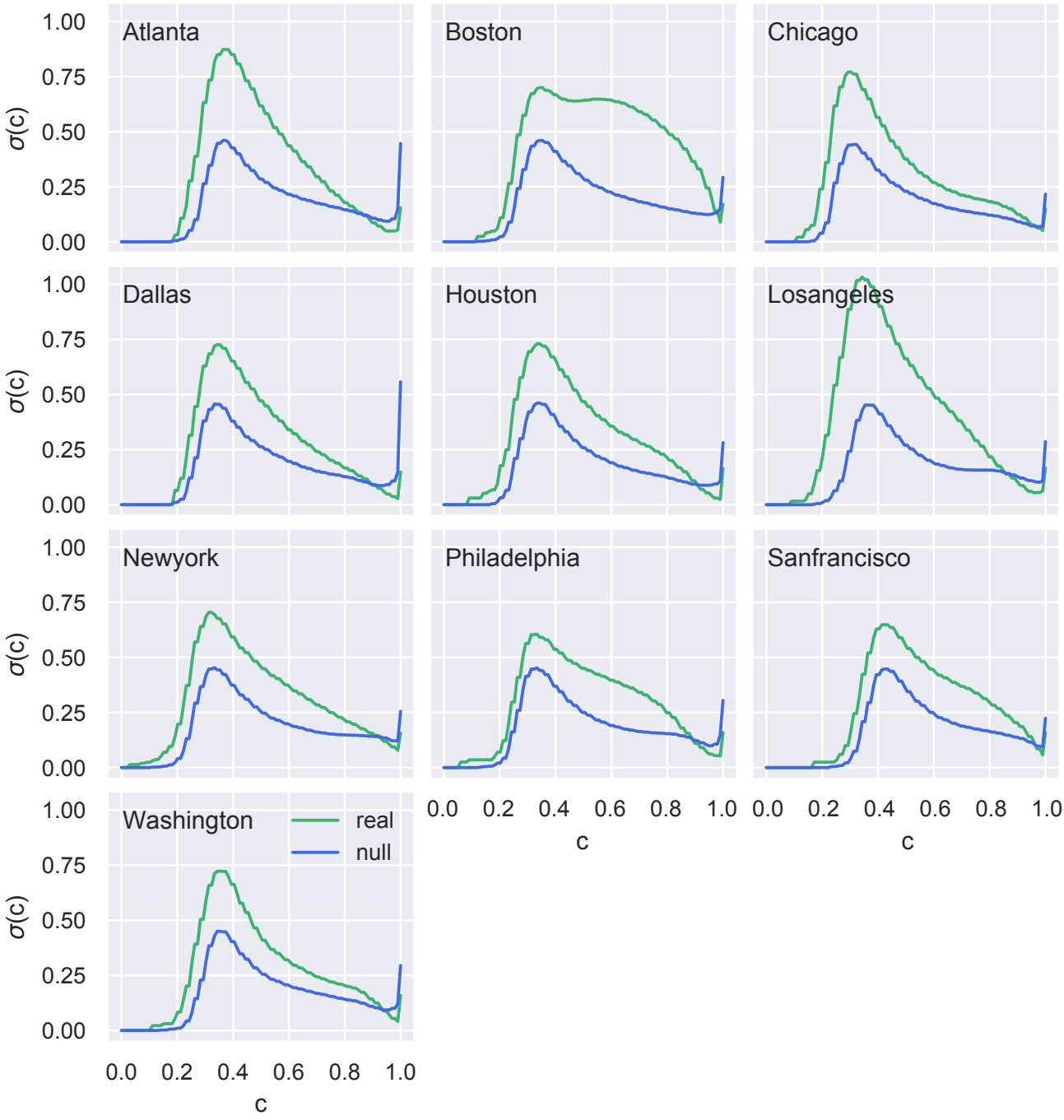

FIG. 8. Coefficient of variance $\sigma(c)$ as a function of fractions c of classes in the real system and the corresponding null model for the metropolitan areas in the US. The values of $\sigma(c)$ are normalised by the respective mean $\mu(c)$. Values for $c = 1$ are kept for demonstration of the spurious effects. The flat line for small values of c is due to the number of classes observed at the initial node i being larger than the fraction c considered.

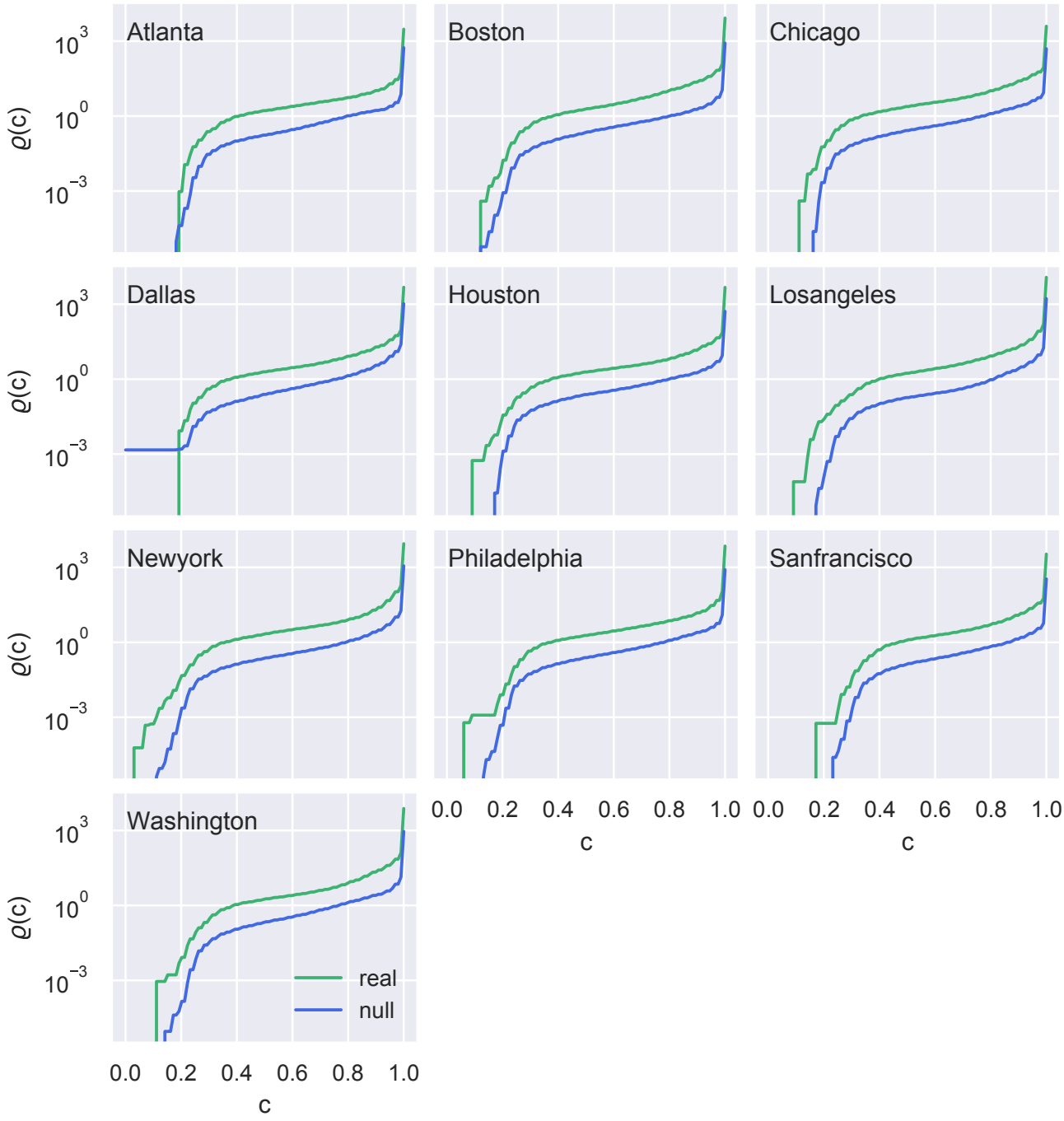

FIG. 9. Spatial diversity $\varrho(c)$ as a function of fractions c of classes in the real system and the corresponding null model for the metropolitan areas in the US.

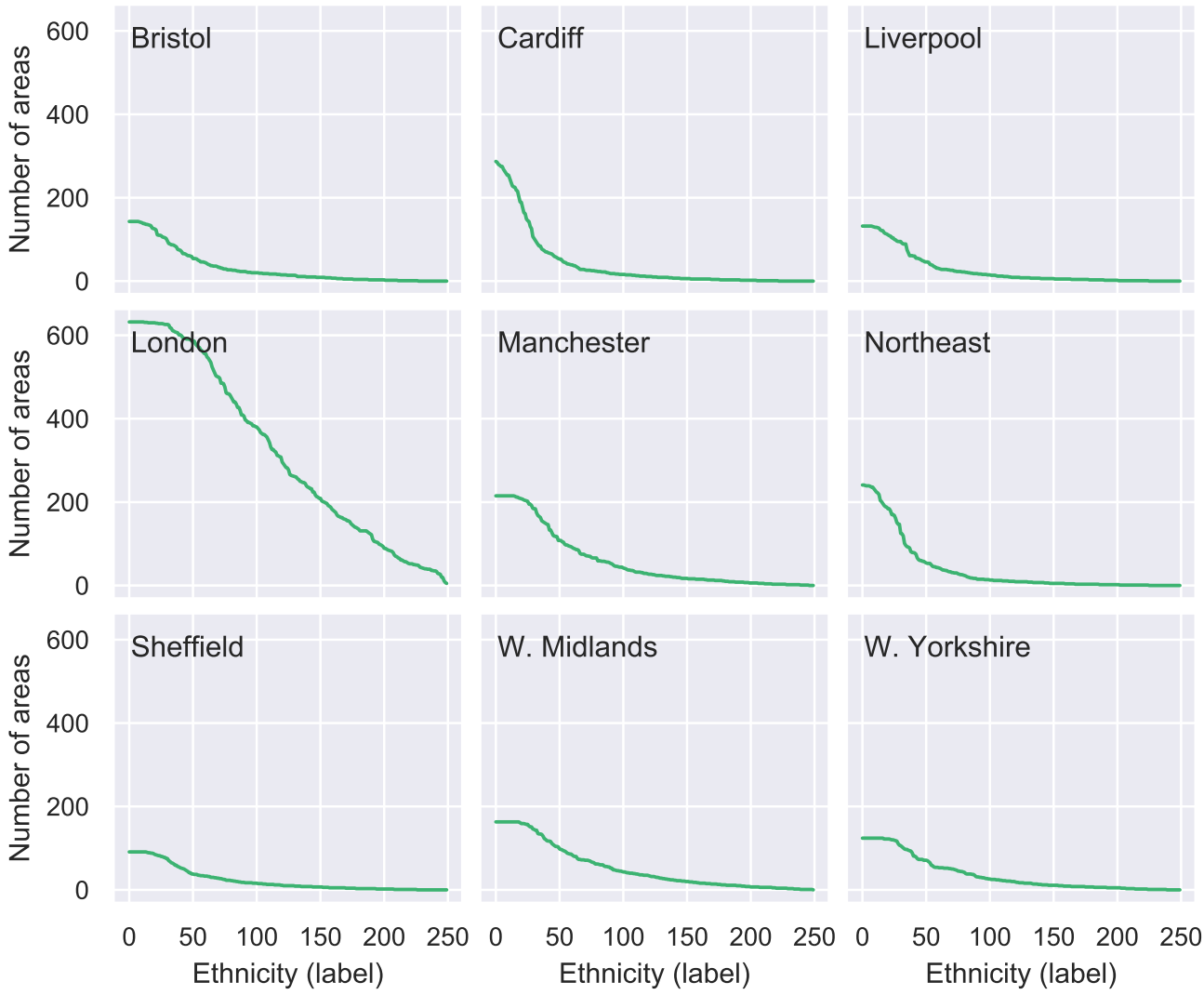

FIG. 10. Number of areal units at which an ethnicity can be found in the metropolitan area. The number labelling the ethnicity corresponds to the column sequence from the UK Census table. Notably, some ethnic groups are present only at a small number of neighbourhoods, which affect the time needed by the random walk to find them.

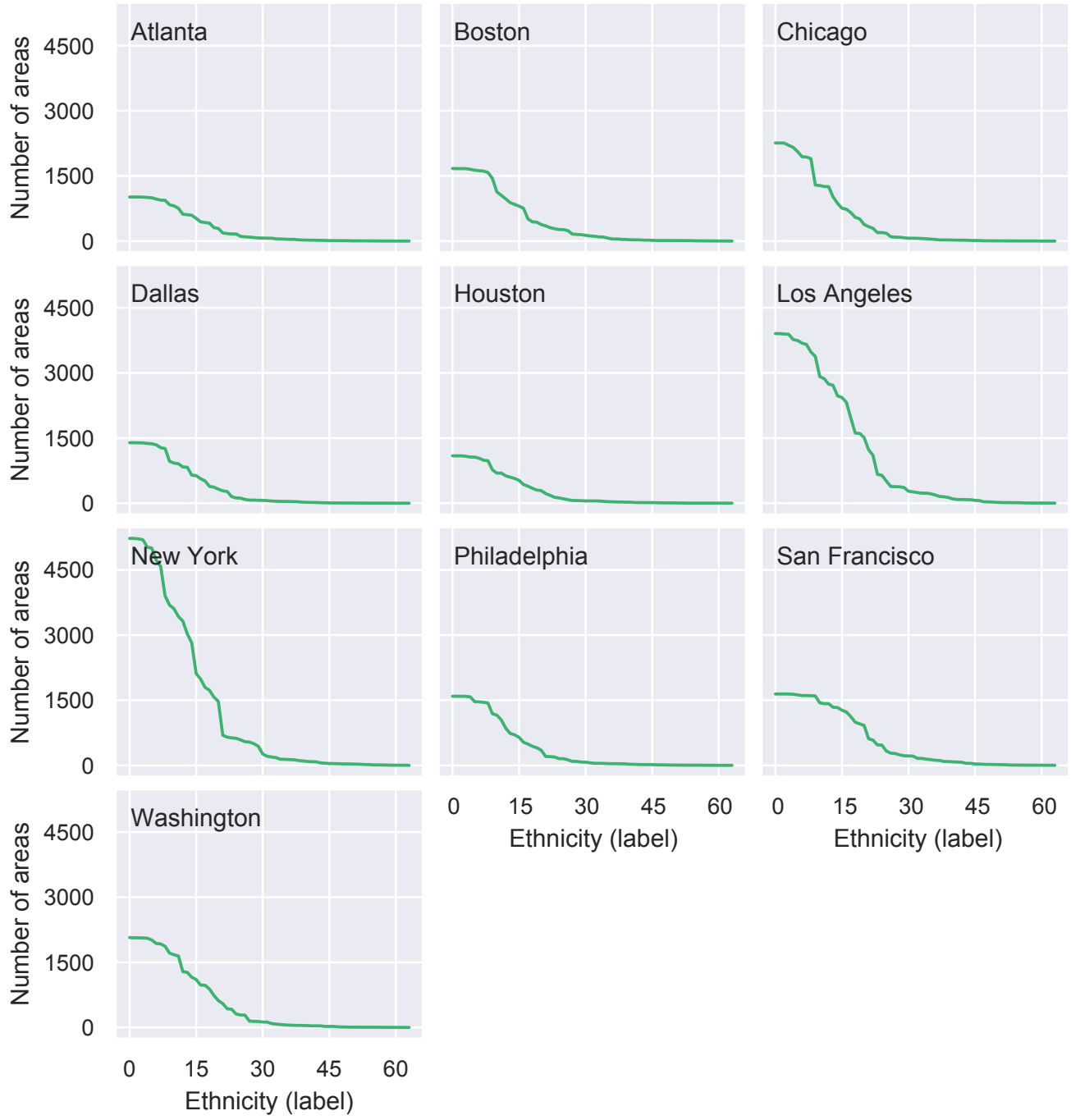

FIG. 11. Number of areal units at which an ethnicity can be found in the metropolitan area. The number labelling the ethnicity corresponds to the column sequence from the UK Census table. Notably, some ethnic groups are present only at a small number of neighbourhoods, which affect the time needed by the random walk to find them.

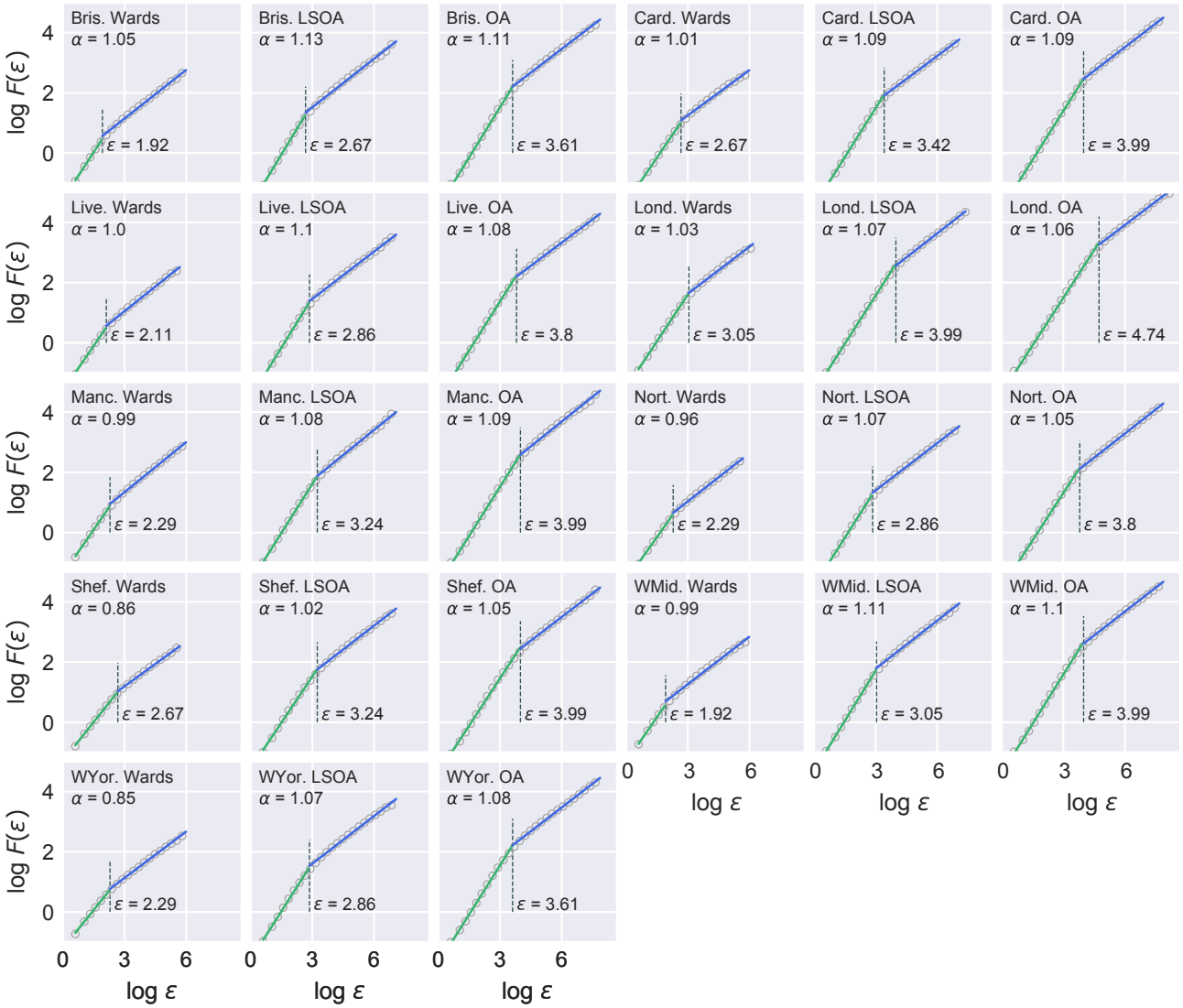

FIG. 12. DFA of the node population entropy reveals two distinct spatial scale regimes across the neighbourhoods of the metropolitan areas in the UK at three different granularities (Wards, Lower Layer Super Output Areas (LSOA) and Output Areas (OA). The value $F(\varepsilon)$ is plotted as a function of ε in log-log scale.

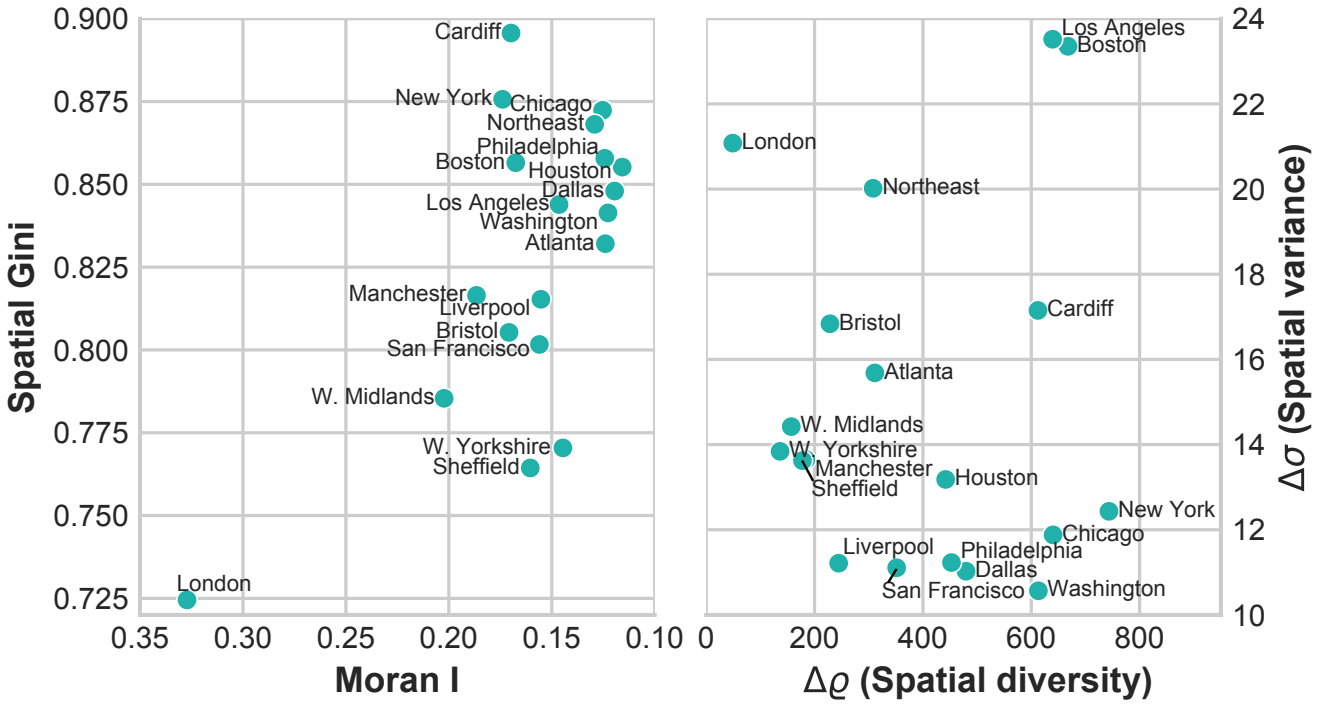

FIG. 13. Quantifying spatial ethnic segregation in urban systems. Alternative view for Fig. 4 in the main text where the results of the Spatial Gini coefficient and Moran I are obtained by calculating the indices for each class in the city and averaging over all classes.

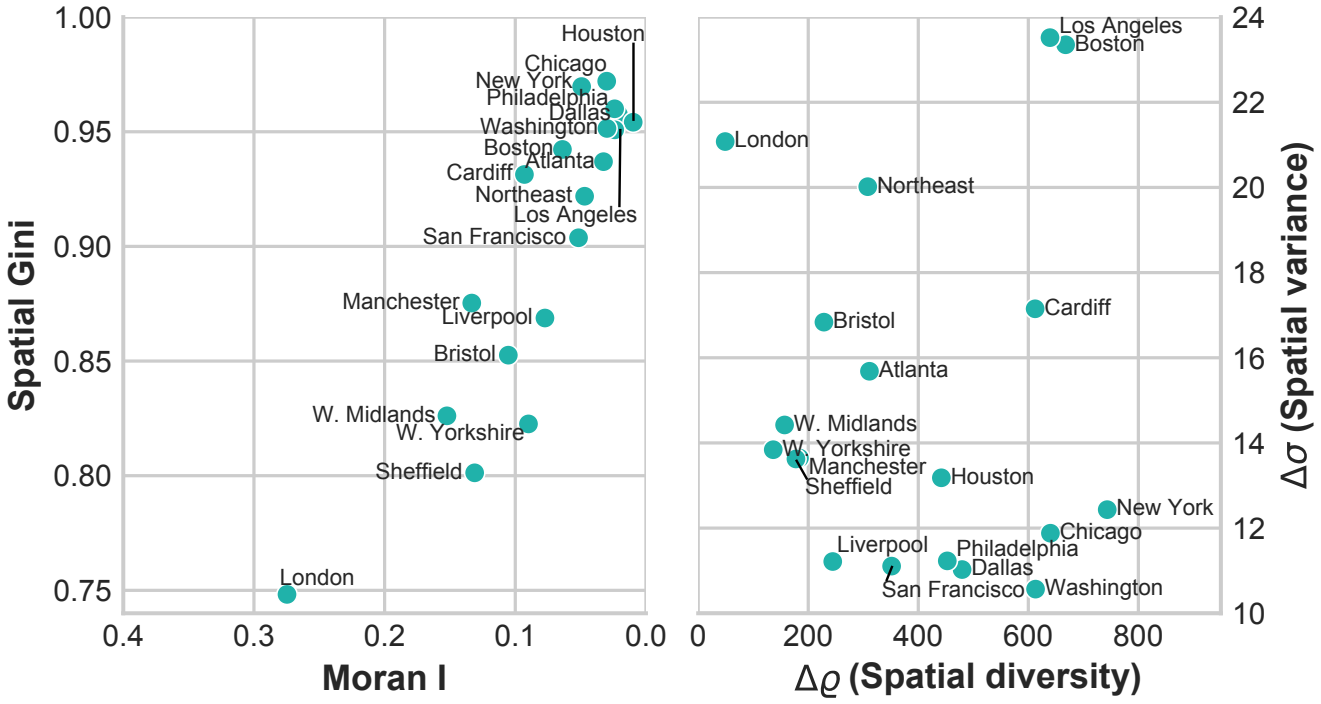

FIG. 14. Quantifying spatial ethnic segregation in urban systems. Alternative view for Fig. 4 in the main text where the results of the Spatial Gini coefficient and Moran I are obtained by calculating the indices for each class in the city and computing the median of all classes.

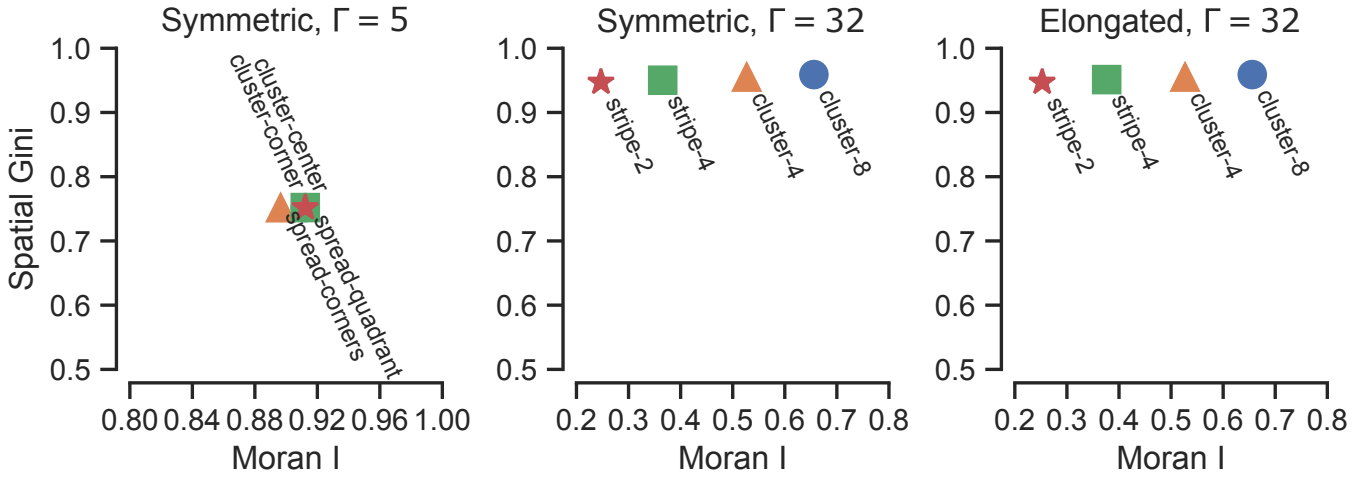

FIG. 15. Quantifying spatial ethnic segregation on synthetic systems. Spatial Gini coefficient and Moran I are reported for the synthetic systems illustrated in the main manuscript. Values are obtained by calculating the indices for each class in the system and computing the median of all classes.

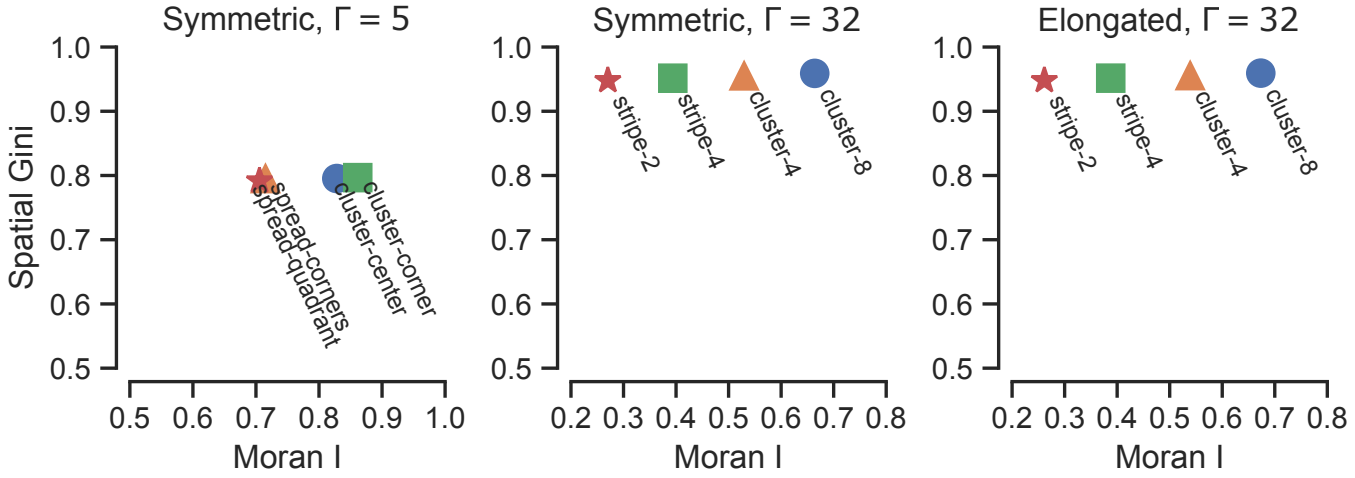

FIG. 16. Quantifying spatial ethnic segregation on synthetic systems. Spatial Gini coefficient and Moran I are reported for the synthetic systems illustrated in the main manuscript. Values are obtained by calculating the indices for each class in the system and computing the average of all classes.