Quantifying ethnic segregation in cities through random walks

Keywords: Spatial segregation, ethnicity, cities, diffusion on networks, random walks

Extended Abstract

Socioeconomic segregation is considered one of the main factors behind the emergence of large-scale inequalities in urban areas and its characterisation is fundamental to understanding a variety of social, technological, and economical processes. A typical example is the segregation of urban areas by socio-economic indicators, including ethnicity, income, education or religion, which is known to be associated with urban wealth, security, and livability [1]. The standard approach in this case is to devise measures of how the local density and heterogeneity of the property under study, as obtained from census data at a given scale, compares with the distribution at the system level, under the assumption that in a non-segregated system, the local distribution of, say, ethnicity would closely mirror the overall distribution in the city level [2]. Despite the vast existing literature on the topic, quantifying spatial segregation is still problematic, mainly because the proposed measures depend on the scale at which neighbourhoods are defined, on the granularity of the census data, or on the presence of one or more free parameters.

In this work, we propose a framework to quantify the multilevel segregation of a city, and to compare the segregation of different cities, based on the statistics of random walks on graphs [3, 4, 5]. The adjacency relations between an urban system's neighbourhoods (census areas) are represented as a spatially-embedded graph $G(V,E)$. We assume that the population of the city is divided into Γ classes (which could correspond to ethnic groups, income or education, etc.), so that each neighbourhood $i \in V$ is assigned to a vector $x_i \in \mathbb{R}^\Gamma$ whose components are the number of people of each class living in i. We consider the time series of node properties generated by the trajectories of an unbiased random walker through the graph G, and we analyse the spatial distribution of the Class Coverage Time (CCT). In particular, we look at the average fraction of distinct classes encountered by the walker up to time t when it started from node i at time 0, namely $\overline{\mathscr{W}}_i(t)$. We define the Class Coverage Time (CCT) of node i at level c as the expected number of steps after which a walker started at i has encountered a fraction c of the Γ classes for the first time, $\underset{t}{\operatorname{argmin}} \left\{\overline{\mathscr{W}}_i(t) \geq c\right\}$. We consider the CCT across all nodes and characterise a city by its mean $\mu(c)$, its coefficient of variation $\sigma(c)$ and the level of local spatial correlation $\rho(c)$. These measures are compared with their corresponding quantity measured in a null-model, where the class distribution x_i is reshuffled across the same graph G through spatial permutations. We refer to the average deviation of the real quantities to their null-model by $\Delta\mu$ "spatial heterogeneity", $\Delta\sigma$ "spatial variance" and $\Delta\rho$ "spatial diversity".

Higher levels of spatial variance $\Delta\sigma \gg 0$ are associated with a more unbalanced spatial distribution of classes, i.e., citizens experience large variations in the time needed to encounter all the other ethnicities depending on where they live. Low levels of local spatial diversity $\Delta\rho$ indicate that there is no significant difference in the coverage time of neighbouring areas. When $\Delta\rho \gg 0$, the differences between neighbouring nodes are substantial, and segregation is influenced by clusters with similar ethnicity distributions. We analyse metropolitan areas in the US and the UK using geo-referenced data from Census where the population is divided into up to 64 and 250 groups, respectively. In Fig. 1b, we look closely at the behaviour of $\mu(c)$, $\sigma(c)$

and $\rho(c)$ in London, which is well-known for being characterised by strong ethnic segregation [6]. Indeed, some areas of the city clearly exhibit substantially larger values of class coverage time as indicated by high $\Delta\sigma$. Panel (a) reports all urban areas analysed here in the $\Delta\mu/\Delta\rho$ and $\Delta\sigma/\Delta\rho$ planes. We note that Boston is placed at the very far ends of both planes. Indeed, the spatial distribution of $C_i(c)/C_i^{\mathrm{null}}(c)$ across Boston (map visible in (c)) shows that the city exhibits opposing spatial patterns. The northern side is characterised by many areas with $C_i(c)$ up to three times larger than in the null model. Los Angeles, which has a similar wide hot spot of segregated areas in the centre, shows comparably high spatial variance but lower levels of spatial heterogeneity.

Remarkably, these measures based purely on diffusion can capture information about the spatial organisation of ethnicities in a city. We find that the distribution of class coverage times provides useful insight into the microscopic, mesoscopic, and macroscopic organisation of ethnicities throughout a city. Furthermore, we find that class coverage times correlate with many deprivation indices more strongly than other classical segregation measures do. These results suggest that measuring multi-scale urban segregation through random walk statistics is potentially more informative than many other current approaches. The framework proposed here is quite flexible and extensible. Although we have mostly focused on the time to find a certain ethnic group, irrespective of its relative abundance, one can refine the analysis by defining the vector x_i according to the local abundance of ethnic groups. Likewise, one can analyse any number of classes, avoiding the many potential biases introduced by aggregating smaller ethnic groups into arbitrary classes. The consistent behaviour of $\Delta\mu$, $\Delta\sigma$ and $\Delta\rho$ across different scales is a desirable property of segregation measures, as also pointed out by classical and more recent works [7, 8]. The fact that these measures are appropriately normalised by comparing with the corresponding null-models makes them suitable for comparing the spatial heterogeneity of the same variable in different systems, irrespective of their peculiar size and shape, of the actual number of different classes or categories available in each system, and of the granularity at which spatial information is aggregated.

References

[1] Barthelemy, M. *The Structure and Dynamics of Cities: Urban Data Analysis and Theoretical Modeling* (Cambridge University Press, Cambridge, 2016).

[2] Reardon, S. F. & O'Sullivan, D. 3. Measures of Spatial Segregation. *Sociological Methodology* **34**, 121–162 (2004).

[3] Noh, J. D. & Rieger, H. Random Walks on Complex Networks. *Physical Review Letters* **92**, 118701 (2004).

[4] Masuda, N., Porter, M. A. & Lambiotte, R. Random walks and diffusion on networks. *Physics Reports* **716-717**, 1–58 (2017).

[5] Ballester, C. & Vorsatz, M. Random walk-based segregation measures. *The Review of Economics and Statistics* **96**, 383–401 (2014).

[6] Barros, J. & Feitosa, F. F. Uneven geographies: Exploring the sensitivity of spatial indices of residential segregation **45**, 1073–1089.

[7] Chodrow, P. S. Structure and information in spatial segregation. *Proceedings of the National Academy of Sciences* **114**, 11591–11596 (2017).

[8] Olteanu, M., Randon-Furling, J. & Clark, W. A. V. Segregation through the multiscalar lens. *Proceedings of the National Academy of Sciences* **116**, 12250–12254 (2019).

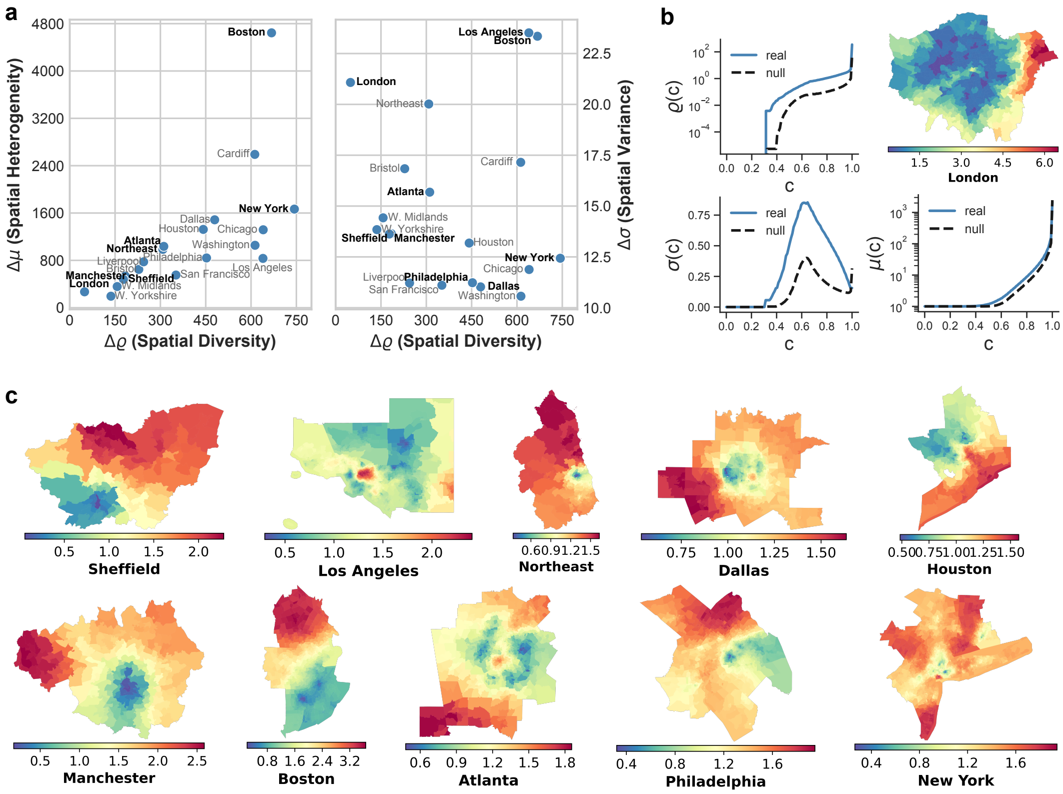

Figure 1: **Class coverage times and ethnic segregation in urban systems**. **(a)** Metropolitan areas in the US and the UK in the $\Delta\mu/\Delta\rho$ and $\Delta\sigma/\Delta\rho$ planes where they are compared by the average deviation from their corresponding null-models. **(b)** Examples of the class coverage time distributions for London where the spatial diversity $\rho(c)$, spatial variance $\sigma(c)$, spatial heterogeneity $\mu(c)$ and their values in the corresponding null-model (black dashed lines) are plotted as a function of the fraction of visited classes c. **(c)** Maps of the metropolitan areas marked in bold in panel (a). The normalised class coverage time $C_i(c)/C_i(c)^{\text{null}}$ for $c = 0.7$ provides detailed insights about the structure of segregation at neighbourhood level.