ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

Spatial analysis in high resolution geo-data

Madalina Olteanu[1,2] and Julien Randon-Furling[2] and William Clark[3]

1- MaIAGE - INRA, Jouy-en-Josas - France

2- SAMM - Université Panthéon Sorbonne, Paris - France

3- Dpt of Geography - University of California, Los Angeles - USA

Abstract. The analysis of spatial dissimilarities across cities often relies on pre-defined areal units, leading to problems of scale, interpretability and cross-comparisons. Furthermore, traditional measures of dissimilarities tend to be single-number indices that fail to capture the complexity of segregation patterns. We present in this paper a method that allows one to extract and analyze information on all scales, at every point in the city, through a stochastic sequential aggregation procedure based on high-resolution data. This method provides insightful visual representations, as well as mathematical characterizations of segregation phenomena.

1 Introduction

Geographical maps of local densities of group populations may exhibit spatial patterns, but often these are blurred by the sheer variety of details when the resolution is high. Indices have been devised to try and capture the details of any regularity emerging in the spatial distributions [1]. The (numerous) existing indices may be classified in at least two categories [2, 3]. On the one hand, zone-based indices such as the dissimilarity index [4, 5, 6], the proximity index [7] or the concentration profile [8] work at fixed scales. They are all liable to the Modifiable Areal Unit Problem (MAUP) [9]. On the other hand, surface-based measures [10, 11] use a continuous population density surface to circumvent the MAUP. But data most often comes as already aggregated units, so these indices usually require refined statistical interpolation techniques. Furthermore, they are not scale-free, since one has to select values for the radius within which the population density is estimated and the dissimilarity indices computed. Another class of indices largely used in spatial statistics is that based on spatial autocorrelations [12] sometimes coupled to a subsequent clustering. Although easy to compute and helpful in practice, these require introducing a dependence structure on the grid of spatial units.

Recently, multiscalar approaches have been proposed [13, 14, 15]. The method we present here is multiscalar, and also both scale-free and non-parametric. It aims at extracting all the information available in the data as scale is varied from the finest possible grain to the whole region of analysis.

2 Method

The first step of our procedure consists in computing a dissimilarity trajectory associated with each spatial unit in the dataset, and encoding the difference

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

between an expanding neighborhood around the starting unit and the whole city. We assume here that the spatial information is available as a square lattice at an already "basic" aggregated level, but other configurations such as geo-localized individual data or aggregated irregular polygons may be similarly dealt with. To each spatial unit $(u_i)_{i=1,\dots,N}$, is associated an empirical distribution of some random variable, measured on the n_i individuals belonging to unit u_i. We then sequentially aggregate around u_i all other spatial units, according to a rule consisting here in first randomly selecting the units situated at a supremum distance equal to 1, then 2, and so on. The aggregation procedure is summarized in Figure 1. At each step k of this procedure, k spatial units have been clustered, including the starting one, and one may compute both the empirical distribution $\hat{f}_{i,1:k}$ of the aggregated population on the k units, as well as the dissimilarity with respect to the distribution of the whole city, $d\left(\hat{f}_{i,1:k}, f_0\right)$. To measure this dissimilarity, we use the Kullback-Leibler (KL) divergence [16].

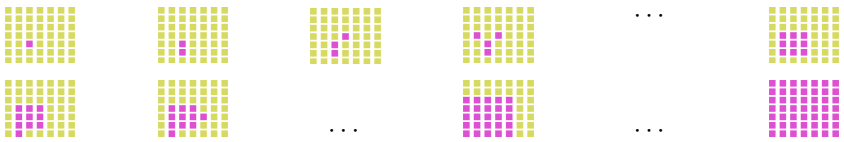

Fig. 1: Aggregation procedure

Once one has aggregated all N spatial units around unit u_i, one obtains that $\hat{f}_{i,1:N} = f_0$ and $d\left(\hat{f}_{i,1:N}, f_0\right) = 0$. Each trajectory naturally converges to the city average and the set of trajectories forms a fingerprint of the city for the variable under consideration. But this convergence is achieved more or less rapidly. If the city were well mixed, then each trajectory would converge in just a few steps. The next stage in our procedure quantifies empirically the speed at which this convergence is achieved individually, starting from any point in the city. For a given spatial unit u_i and for the associated KL-divergence trajectory $\left(n_{i,1:k}, d\left(\hat{f}_{i,1:k}, f_0\right)\right)_{k=1,\dots,N}$, where $n_{i,1:k}$ is the size of the aggregated population on the first k units around u_i, we compute the convergence "time" to the city. We define it, for any threshold $\delta \geq 0$, as the aggregated population size for which the KL-divergence trajectory enters (and remains) within the interval $[0, \delta]$:

$$\tau_{i,\delta} = \min_{k=1,\dots,N} \left\{ n_{i,1:k} \mid d\left(\hat{f}_{i,1:\tilde{k}}, f_0\right) \leq \delta, \forall \tilde{k} \geq k \right\}$$

Furthermore, we remove the arbitrariness induced by selecting an a priori threshold δ by integrating the convergence times, on all possible values of the threshold (the upper bound $\delta_{i,\max}$ is necessarily finite and is equal to the maximum value of the KL-divergence trajectory):

$$\Delta_i = \int_0^{\delta_{i,\max}} \tau_{i,\delta} d\delta \ .$$

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

Eventually, our procedure produces a coefficient Δ_i for each spatial unit u_i, which encompasses the level of *distortion* in the image of the city perceived from the corresponding spatial unit.

Defined as above, distortion coefficients depend on the individual values of the KL-divergence trajectories as well as on the number of inhabitants in the city. Hence the need for a normalization constant, which will make inter-city comparisons and inter-variable analyses possible. From geographical analysis and information theory, we argue that the normalization constant should be taken equal to the distortion coefficient computed on the theoretical spatial configuration which maximizes segregation. In the case of a Bernoulli variable, with a proportion $p_0 < 0.5$ of group A, the trajectory with the maximum distortion coefficient is that consisting in first aggregating exclusively individuals of type A and then all the individuals of type B. This normalization constant, $\tilde{\Delta}$, depends in this case on p_0 only and may be explicitly computed as:

$$\tilde{\Delta} = -p_0 \log(p_0) + \int_{p_0}^{1} \left[\frac{p_0}{x} \log\left(\frac{1}{x}\right) + \frac{x - p_0}{x} \log\left(\frac{x - p_0}{x(1 - p_0)}\right) \right] dx \ .$$

We illustrate next the proposed procedure by studying the spatial distributions of foreign-born inhabitants in the city of Paris.

3 Data and results

The data we use comes from the D4I challenge on the integration of migrants in cities launched by the Joint Research Center of the European Commission (`https://bluehub.jrc.ec.europa.eu/datachallenge/`). It is a large, high-resolution dataset with counts of foreign-born inhabitants for each 100x100m cell on a regular grid.

	EU27	Non-EU	Chinese	Algerian
Migrant population	92,026	240,307	27,645	30,126
% of the entire city population	4.09%	10.69%	1.23%	1.34%
Max. theor. dist. coeff. $\tilde{\Delta}$	0.269	0.404	0.138	0.146

Table 1: Proportions of certain migrant communities in the city of Paris

We applied our method on the entire commuting zone of Paris, and also on three other European capitals (Rome, Madrid, Berlin).We present here the Paris results on four communities – two very general (EU27 migrants and non-EU migrants), and two specific (Chinese and Algerian). Paris, with its 2,248,435 inhabitants, was divided in 9,156 spatial units. The proportions of each of the communities, as well as their maximum theoretical distortion coefficient, used hereafter as normalization constant, are summarized in Table 1. The local densities per spatial unit are represented in Figure 2. As informative as they may be, these maps do not provide any clear picture of the spatial patterns for each community. Neither do they easily allow for comparisons. On the upper maps,

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

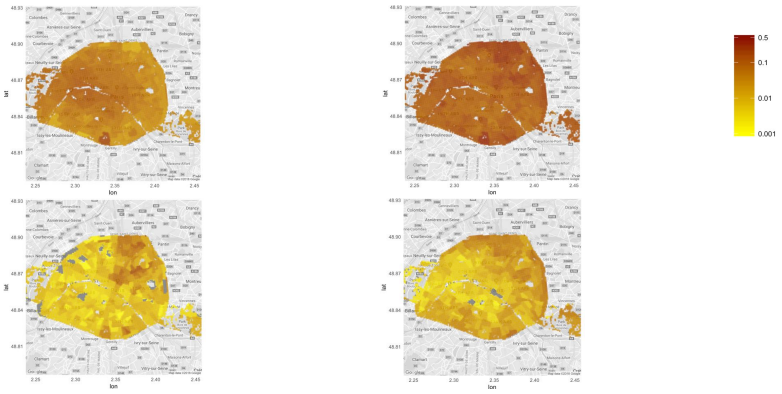

Fig. 2: Local densities per spatial unit in logarithmic color scale (upper left - EU migrants; upper right - non-EU migrants; lower left - Chinese-origin migrants; lower right - Algerian-origin migrants). Grey areas correspond to densities with less than 0.1% of the selected community.

one may easily spot the international Cité Universitaire in the south of the city, almost exclusively inhabited by foreign students, as well as the northern districts, where the presence of non-EU migrants seems to be more significant. The Chinese-origin migrants are mainly located in some of the *Rive-Droite* neighborhoods, as well as the south-east of the city, in the 13th district which is sometimes termed "Paris' Chinatown". As for the Algerian migrants, they are mostly distributed in the northern, eastern and southern parts of the city, and this spatial distribution is highly correlated with that of social housing and income [15]. Furthermore, one may also notice that although these two communities are rather similar in terms of global rates, each of them representing about 1.25-1.3% of the entire population, the patterns of their spatial distributions appear to be quite different, as we shall confirm next.

For each of the four communities, we computed the distortion coefficients, and then normalized them with respect to the corresponding theoretical maximum of segregation, $\tilde{\Delta}$. The resulting distributions are plotted in Figure 3. First, let us remark that the distributions of the EU and non-EU migrants are much less dispersed, and with smaller mean values than those of the Chinese and Algerian migrants. This is due to a smoothing effect introduced by the aggregation of many various origins, whereas the patterns of installation appear to be particularly dependent on the country of origin. Second, the distributions of the distortion coefficients for the Chinese and Algerian migrants have larger means and larger dispersions, and also bimodal densities, which suggest at least two categories of spatial units: some with low distortion, from where the correct "perception of the city" is rapidly achieved, and some with high distortion, hence more segregated. This seems to be particularly the case for the Chinese migrants distribution, which also has a heavy right tail, which implies

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0.
Available from http://www.i6doc.com/en/.

the existence of some units, called "hot spots", particularly segregated.

Finally, the normalized distortion coefficients are mapped in Figure 4 in logarithmic scale. Grey and blue areas correspond to low values of distortion, hence to neighborhoods of the city from which the perception is roughly correct on any scale, from a relatively short one. Red areas have high distortion, hence are more segregated than the others. We see for instance that the North-North-Eastern neighbourhoods are more segregated, for Chinese migrants, than Paris' Chinatown. This fact has been totally unnoticed through other indices so far. This is a typical finding of our method.

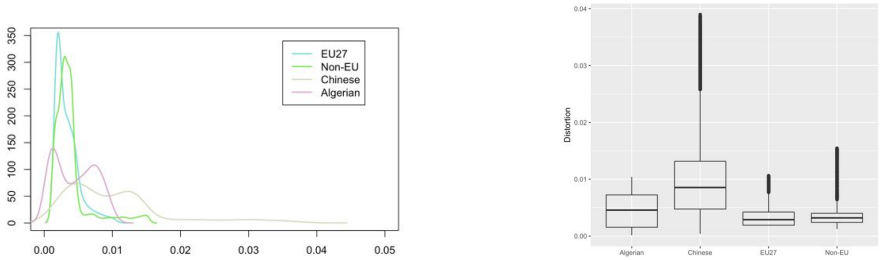

Fig. 3: Normalized distortion coefficients distribution for the four communities (left: estimated densities; right: boxplots).

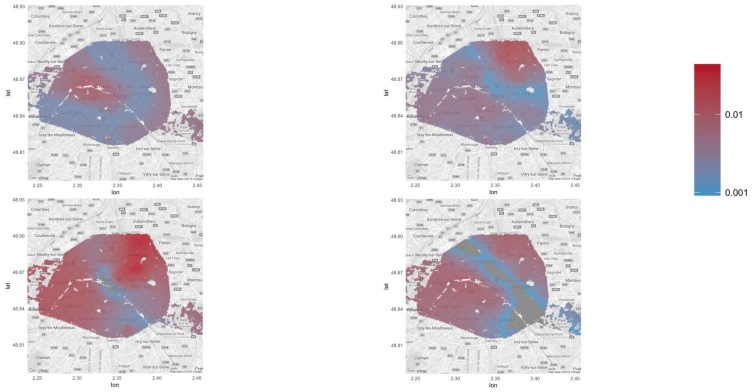

Fig. 4: Normalized distortion coefficients per spatial unit in logarithmic color scale (upper left - EU migrants; upper right - non-EU migrants; lower left - Chinese-origin migrants; lower right - Algerian-origin migrants). Grey areas correspond to coefficients less than 0.001.

4 Conclusion and perspectives

The method presented here provides a simple and powerful tool for visualizing spatial segregation throughout cities and for any variable of interest.By con-

ESANN 2019 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2019, i6doc.com publ., ISBN 978-287-587-065-0. Available from http://www.i6doc.com/en/.

struction, it is a scale-free algorithm, which is a step forward in the analysis of spatial information. This method would also be of great interest on individual geo-localized data, should such data become available – although scaling up to a few dozens/hundreds of millions units would not be completely straightforward. Individual-level data would also allow us to model trajectories by random walks, which would provide theoretical results on first passage times, sojourn times and statistical properties of the distortion coefficients.

References

[1] S. F Reardon and G. Firebaugh. Measures of Multigroup Segregation. *Sociological Methodology*, 32(1):33–67, 2002.

[2] M. R. Kramer, H. L. Cooper, C. D. Drews-Botsch, L. A. Waller, and C. R. Hogue. Do measures matter? comparing surface-density-derived and census-tract-derived measures of racial residential segregation. *International Journal of Health Geographics*, 9(1):29, Jun 2010.

[3] S. Hong, D. O'Sullivan, and Y. Sadahiro. Implementing spatial segregation measures in R. *PLOS ONE*, 9:1–18, 11 2014.

[4] O. D. Duncan and B. Duncan. A methodological analysis of segregation indexes. *American sociological review*, 20(2):210–217, 1955.

[5] D. WS Wong. Spatial indices of segregation. *Urban studies*, 30(3):559–572, 1993.

[6] R. L Morrill. On the measure of geographic segregation. In *Geography research forum*, volume 11, pages 25–36, 2016.

[7] M. J White. The measurement of spatial segregation. *American journal of sociology*, 88(5):1008–1018, 1983.

[8] M. Poulsen, R. Johnson, and J. Forrest. Plural cities and ethnic enclaves: introducing a measurement procedure for comparative study. *International Journal of Urban and Regional Research*, 26(2):229–243, 2002.

[9] S. Openshaw. *The modifiable areal unit problem*. University of East Anglia, 1984.

[10] S. F. Reardon, S. A. Matthews, D. O'Sullivan, B. A. Lee, G. Firebaugh, C. R. Farrell, and K. Bischoff. The geographic scale of Metropolitan racial segregation. *Demography*, 45(3):489–514, Aug 2008.

[11] F.F Feitosa, G. Camara, A. M. V. Monteiro, T. Koschitzki, and M. PS Silva. Global and local spatial indices of urban segregation. *International Journal of Geographical Information Science*, 21(3):299–323, 2007.

[12] P. A. P. Moran. Notes on continuous stochastic phenomena. *Biometrika*, 37(1-2):17–23, 1950.

[13] C. Fowler. Segregation as a multiscalar phenomenon and its implications for neighborhood-scale research: the case of south seattle 1990–2010. *Urban geography*, 37(1):1–25, 2016.

[14] W. A V Clark, E. Andersson, J. Östh, and B. Malmberg. A multiscalar analysis of neighborhood composition in Los Angeles, 2000–2010: A location-based approach to segregation and diversity. *Annals of the Association of American Geographers*, 105(6):1260–1284, 2015.

[15] J. Randon-Furling, M. Olteanu, and A. Lucquiaud. From urban segregation to spatial structure detection. *Environment and Planning B: Urban Analytics and City Science*, page 2399808318797129, 2018.

[16] S. Kullback and R. A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.