When the *k*-spatial entropy is fast and faster using a zoning system

Didier G. Leibovici[1], Konstantinos Daras[1]

[1]Univerisity of Leeds, School of Geography, U.K.
d.g.leibovici@leeds.ac.uk, k.daras@leeds.ac.uk

KEYWORDS: spatial entropy, co-occurrences, geocomputation, zoning system, R

1. Introduction

When a method is computationally intensive one focuses on the theoretical merit forgetting the issue! So far no fast algorithm were found for the *k*-spatial entropy exposed in GISRUK 2010 (Leibovici *et al*. 2010, 2011ab). Whilst packaging it into R (Leibovici 2013), some interesting outcomes surprised me in terms of conceptual aspects, algorithms performances and R coding. This paper addresses the particular method that is both conceptually interesting and very fast to compute.

Extending the *k*-spatial entropy framework to any type of geometries, as for areal data (compositional data, census data), using weights attached to the geometries, and relaxing the one-variable category per geometry, became computationally intensive. The self-*k*-spatial entropy index can overcome partially these issues. With census data, it is desirable to report back the entropic variations within and across a particular zoning system at which the decision-making has a direct impact. Constraining the *k*-spatial entropy methodology approach using a particular zoning, say Wards when the data is collected or simulated at OA levels, keeps the computational time down and can be easily parallelised. Besides using existing zoning systems for describing population dynamics, the proposed approach can also be used for zoning system optimisation (Daras, 2006, Haynes et al. 2007). This last aspect is discussed for two related zoning optimisation:

- finding a zoning system with minimum *k*-spatial entropy across the zones and maximum *k*-spatial entropy within each zone (minAmaxW)
- finding a zoning system with maximum *k*-spatial entropy across the zones and minimum *k*-spatial entropy within (maxAminW).

The minAmaxW optimisation produces a zoning system that differentiates the most a set of most homogeneous zones whilst maxAminW looks for a set of regions the most similar but themselves being the least uniform. Examples using the MoSes data (Birkin et al. 2009), a microsimulation of the evolution of the population in Leeds between 2001and 2031, are shown. The social grade variable (A/B: professional middle managers, C1: other non-manual workers, C2: skilled manual workers, D: other manual workers, and E: on benefit/unemployed) is chosen for this paper.

2. Zoning and entropy

For a set of regions R and the distribution of a categorical variable C over the regions: a set of proportions p_{cr} with $\sum_{c,r} p_{cr} = 1$ representing the distribution of cases by category and by regions, $p_{cr} = n_{cr}/N$, with N as the total population count, one can use the property of the conditional entropy to get:

$$
\begin{aligned}
&H(C,R) = -\sum_{c,r} p_{cr} \log(p_{cr}) \\
&= -\sum_r p_{.r}\, log\, p_{.r} - \sum_r p_{.r} (\sum_c p_{c/r}\, log(p_{c/r}) \\
&= H(R) + H(C/R) = \; H(C) + H(R/C)
\end{aligned} \tag{1}
$$

where $p_{c/r} = p_{cr}/p_{.r}$ with $p_{.r} = \sum_c p_{cr}$ is the conditional probability of the category c from the categorical variable C given the region $R = r$. In other words (1) often called the entropy decomposition theorem (Theil 1972) insures that the entropy of a categorical variable disaggregated over a spatial support is the sum of the entropy of the spatial support margins and the conditional entropy of the categorical variable given the spatial support. The conditional entropy measures the variables association:

$$
0 \leq H(C/R) \leq H(C) \tag{2}
$$

reaching the lower bound when C is completely determined by R and the upper bound when C and R are two independent random variables. Finding a spatial support that tries to reach either bound can be of interest in population sciences. A zoning system explaining most of the categorical variables distributions can facilitate policy implementations but working with a zoning system independent of the studied variables facilitates global policy-making expecting to impact equally in each area.

The data is usually available already at a given scale following a particular zoning system, then with an extra zoning system Z aggregating the given one (1) becomes:

$$
\begin{aligned}
&H(C,R) = H(C,R(Z)) = -\sum_{c,r(z)} p_{cr(z)} \log(p_{cr(z)}) \\
&= -\sum_z p_{..(z)}\, log\, p_{..(z)} - \sum_z p_{..(z)} (\sum_{c,r \in z} p_{cr/(z)}\, log(p_{(cr/(z)}) \\
&= H(Z) + H((C,R)/Z) = \; H(Z) + H(R/Z) \; + H(C/(R,Z))
\end{aligned} \tag{3}
$$

where Z is seen as a nesting factor in relation to R. The term $H((C,R)/Z)$, the conditional entropy of the set of all local observations at R level given the Z aggregated level is the expectation of the local entropies, *i.e.* over all the r regions within the z locally aggregating. Therefore the decomposition in (3) has the advantage of allowing a map representation at the Z aggregating scale whilst also preserving a global measure assessment (the decomposition itself).

3. Self-*k*-spatial entropy

The Shannon entropy on the spatially distributed values of C, the joint distribution of the categorical variable and the support R, has some drawback. A permutation of the regions will give the same entropy value, as well as (1). This is not reflecting the spatial pattern of their occurrences. Leibovici (2009) introduced a spatial entropy index using directly the collocation of the occurrences, building a distribution related to the clustering of the observations: the *k*-spatial entropy. A collocation is defined by vicinity, *e.g.* a maximum distance between k occurrences (k being the number of events to be considered in one collocation: pairs, triples, or quadruples). The co-occurrence distribution can be multivariate, the running index for the distribution c_{oo}referring to all combination of the categories.

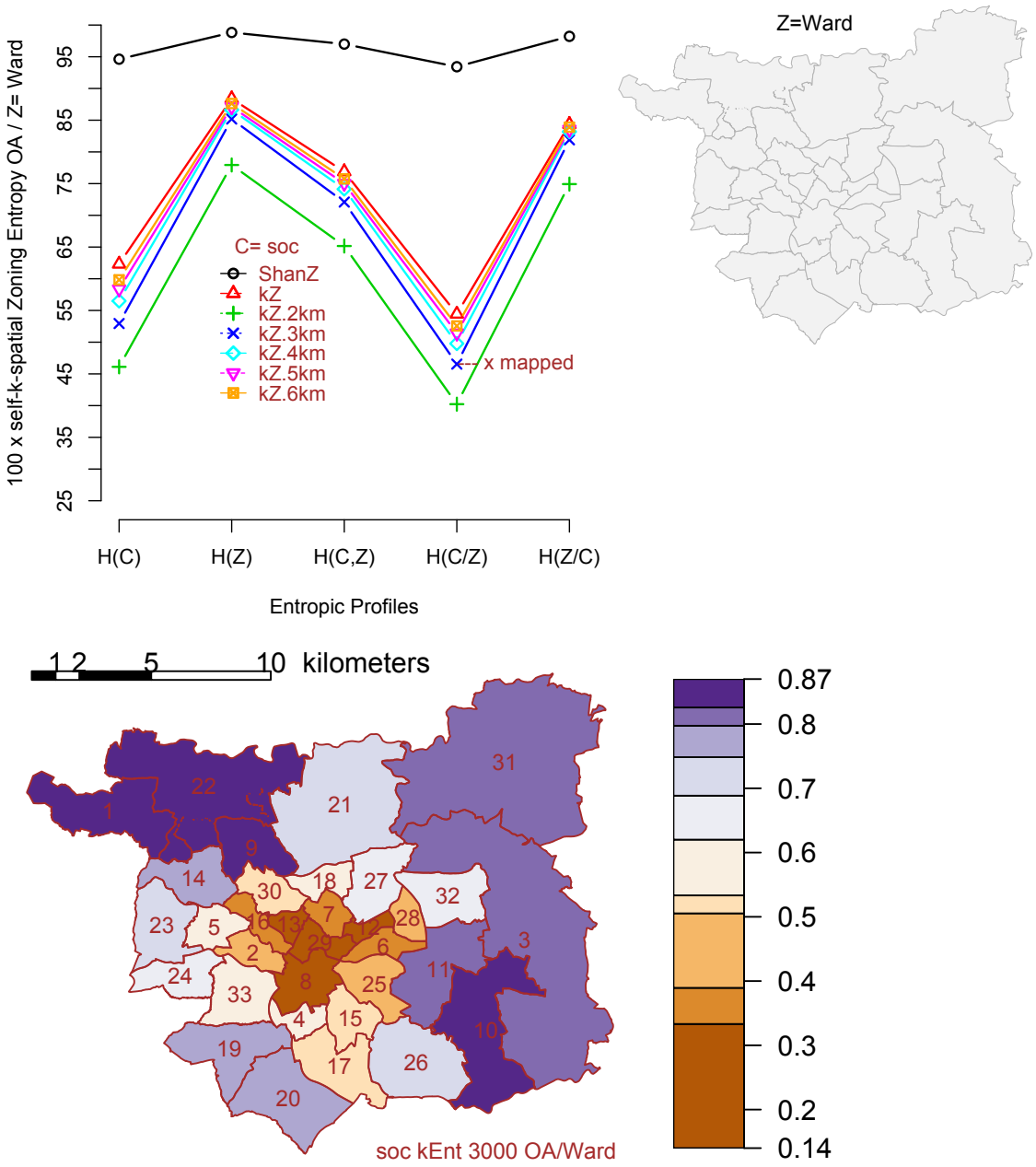

Figure 1. Entropy decompositions with Ward zoning (top panel): all for the Shannon and the self-*k*-spatial zoning entropy (bottom panel): local values of the within self-*k*-spatial entropies at 3000m

Leibovici (2011b) introduced a simpler index, the self-*k*-spatial entropy (4), looking only at co-occurrences of one category with itself: $p_{c_{oo},d} = p_{iii,d}$ for example with $k = 3$, so only the hyper-diagonal of the co-occurrence table is used:

$$H^{s}_{kS}(C,d) = -{}^{1}/_{\log(n_C)} \sum_c p_{cc\ldots c,d} \log(p_{cc\ldots c,d}) \qquad (4)$$

The classical entropy is derived from the distribution of the contingencies of each category whilst the self-*k*-spatial entropy is derived from the spatial co-occurrences of each category.

The vicinity defined by the collocation distance d can be modified to take into account a zoning system Z. The modified statistics (the self-*k*-spatial zoning entropy) (5) still counts

the co-occurrences using the collocation distance d but only within each zone z which acts as a boundary limit, as for the within Z entropies in equation (3):

$$H^s_{ZkS}(C,d) = H^s_{kS}(C,\{Z,d\}) = -{}^{1}\!/_{\log(n_C)} \sum_c p_{cc\ldots c,\{Z,d\}} \log(p_{cc\ldots c,\{Z,d\}}) \quad (5)$$

When the distance d become large enough (5) corresponds to a $H^s_{kS}(C,\{Z\})$, the simplest approach where the vicinity of the co-occurrences is fixed and is the same for all observations within the same zone z. Considering the contingencies of co-occurrences as aggregated or disaggregated in the same way as for contingencies of occurrences, the decomposition (3) is valid for this entropy (re-weighted by the factors of $\log(n_m)$).

Figure 1 illustrates this decomposition for the social grades variable. In (5) the subdivision R disappears as the extension to areal data uses a weighted geometry approach (allowing all the observations to contribute as if we had point data), therefore the decomposition used is (1), is at the coarser scale Z but with areal co-occurrences made at the finer scale R.

4. Zoning optimisation

The generic zoning optimisation approach (Daras 2006) aggregated the 2439 OAs into 33 zones (as many as Wards) using a homogeneity objective function (6), for a single attribute variable, here the within variance for the grouping into N_Z zones:

$$Z_{opt} = \underset{\substack{Z \\ Z \succ R \\ Z \in \mathcal{Z}}}{arg\min}\ t(y)(Id - P_Z)y\,/(n - N_Z) \quad (6)$$

where the numerator is just expressing using projectors the sum of squares of residuals from the local mean for each zone of the attribute y, and Z aggregating R, $(Z \succ R)$ belonging to a range of valid zoning $\mathcal{Z}$ defined by a set of constraints such as the compactness of the shapes (the mean of the distances to the centroid). As no initial threshold is chosen for the compactness, the constraint is operating in a competing way during the algorithm: moving a unit to one zone or another.

The attribute chosen is the proportion of manual workers (aggregating social grades C2 and D). Figure 2 shows clearly the improvement from Figure 1 concerning the overall conditional entropy pursuing a maxAminW objective but the entropic profile is flatter, nonetheless lower than Figure 1. As the new zones have now homogeneous proportions of manual workers the differences between Figure 1 and 2 is due to the co-occurrences of other social grades knowing also that collocation of different categories will have transformed the profiles of the new zones. For example part of the lower entropies in the center on Figure 1 is due to a high proportion of E relatively to A/B or C1, which is transformed on Figure 2 to get about the same and reversed for the areas overlapping wards 31 and 21.

The self-*k*-spatial entropy has an easier interpretation and is faster to compute than the multivariate *k*-spatial entropy. Using a zoning system to constraint the computation has the advantage of being even faster but also to provide a localisation of the patterns. Nonetheless the choice of Z is of concern. Because of the decomposition (1) or (3), the minAmaxW optimisation influence also the whole entropy:

$$Z_{minmax} = \underset{\substack{Z \\ Z \succ R \\ Z \in \mathcal{Z}}}{arg\min}\ \alpha H^s_{ZkS}(Z) + \beta / H^s_{ZkS}(C/Z) \quad (7)$$

where the optimisation choices: $(\alpha + \beta) = 1$, allow emphasizing more or less on one term in the optimisation and within a set of quality constraints fixed by the ensemble $\mathcal{Z}$ such as number of zones, minimum number of population etc. The maxAminW is derived in a similar way to obtain Z_{maxmin}.

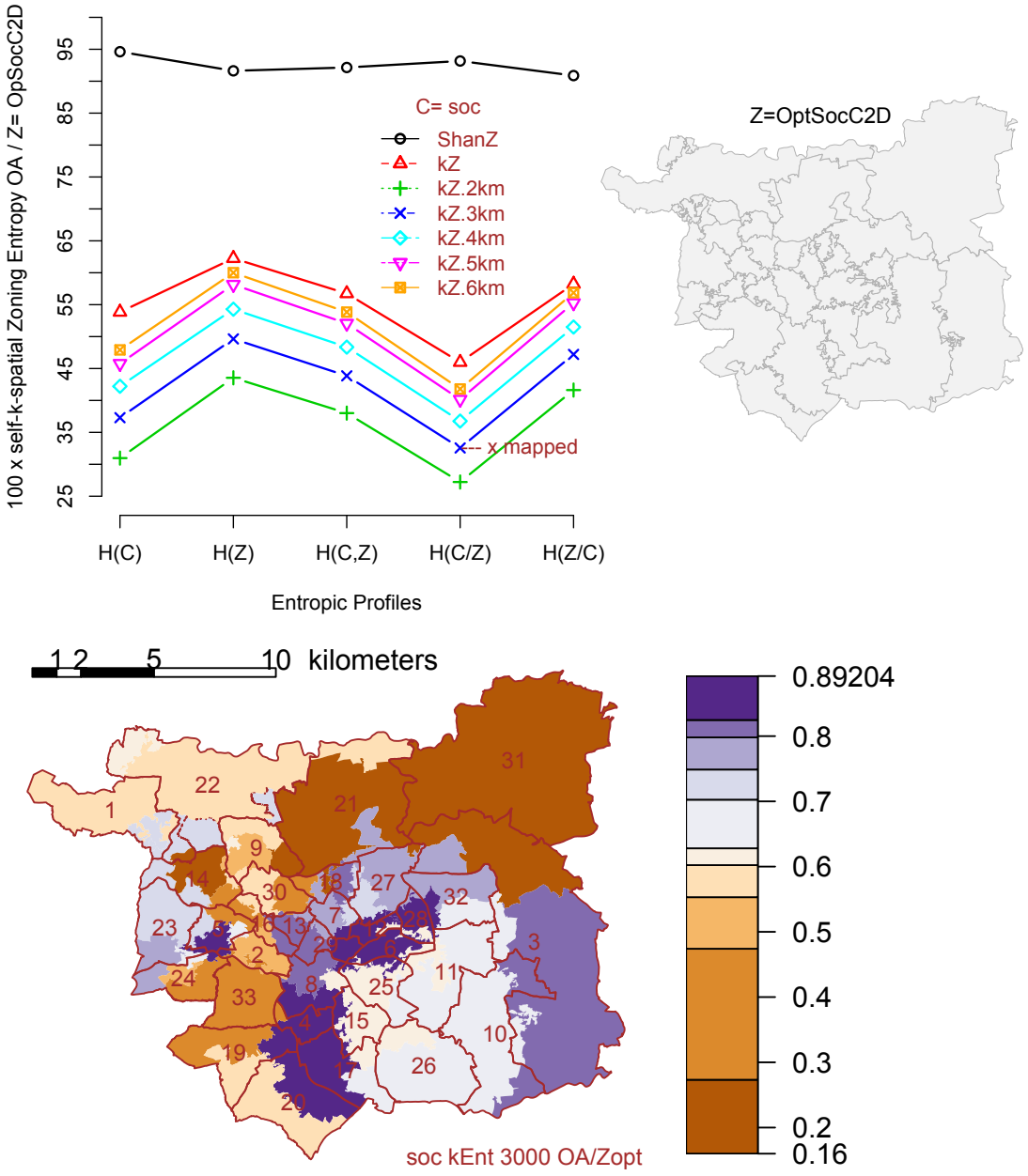

Figure 2. Entropy decompositions with the optimised zoning (top panel): all for the Shannon and the self-*k*-spatial zoning entropy. (bottom panel): local values of the within self-*k*-spatial entropies at 3000m (with Wards overlaid)

The social grade zoning optimisation with the minAmaxW and maxAminW is meant to improe any independent zoning approach, as on Figure 2, towards the given objective. This result with other examples for other variables will be shown for the MoSes data.

5. Discussion

Using a zoning system speeds up considerably the computational time for the self-*k*-spatial entropy, itself faster than the multivariate *k*-spatial entropy. Nonetheless the sum over each zone of Z of the co-occurrence counts is only an approximation of the whole area co-occurrence distribution, as cross-zone border counts are discarded. The value of the self-*k*-spatial entropy at collocation distance of 3000m is 0.358 so quite close to 0.529, the H(C) value on Figure 1 but closer to 0.373, the H(C) value on Figure 2. Removing cross-border co-occurrences have here a global "smoothing" effect (for both zonings) which is recovered locally within the zoning. The is compatible finding a zoning which reveals the spatial patterns associated with the categorical variable, whilst loosing on the re-aggreagated statistic. The application for zoning optimisation opens up the choices of criteria for this type of spatial clustering where homogeneity and heterogeneity integrates a spatial component: the distribution of co-occurrences.

6. Acknowledgements

This work has been funded by the NSRC Tasliman (geosTapial datA anaLysIS and siMulAtioN) project www.geotalisman.org

References

Birkin M.H Townend P Turner A Wu B.M and Xu J (2009) MoSeS: A Grid-enabled spatial decision support system. *Social Science Computing Review*, **27(4)** pp493-508

Daras K (2006) *An information statistics approach to zone design in the geography of health outcomes and provision*. Ph.D. Thesis, University of Newcastle, UK pp 206

Haynes R Daras K Reading R and Jones A (2007) Modifiable neighbourhood units, zone design and residents' perceptions. *Health & Place* **(13),** pp812–825

Leibovici D.G (2009) Defining Spatial Entropy from Multivariate Distributions of Co-Occurrences. *COSIT'09 Conference On Spatial Information Theory*, Aber Wrac'h, France, September 21-25, 2009, *Lecture Notes in Computer Sciences*, **(5756/2009),** pp392-404

Leibovici D.G Bastin L Anand S Swan J Hobona G and Jackson M (2010) Spatially Clustered Associations in Health GIS "mashups". *18th Annual Conference GISRUK*, April 2010. University College London, London, UK

Leibovici D.G Bastin L Anand S Hobona G and Jackson M (2011a) Spatially Clustered Associations in Health related geospatial data. *Transactions in GIS*, **15(3)** pp347-364

Leibovici D.G Bastin L and Jackson M (2011b) Higher-Order Co-occurrences for Exploratory Point Pattern Analysis and Decision Tree Clustering on Spatial Data. *Computers & Geosciences*, **37(3)** pp382-389

Leibovici D.G (2013) A *k*-spatial entropy framework R- package: *k*OO. (to be submitted to

Theil H (1972) *Statistical Decomposition Analysis*. Amsterdam: North Holland.

Wu B.M and Birkin M.H (2012) Agent-Based Extensions to a Spatial Microsimulation Model of Demogaphic Change. *In:* A.J. Heppenstall et al. (eds) *Agent-Based Models of Grographical Systems,* Springer Sciences Business Media, pp347-360

Biography.

Dr. Didier G. Leibovici is a research fellow in geocomputational statistics combining his background, as a statistician and geomatician in epidemiological research and landscape changes in agro-ecology, with geospatial modelling. Besides these spatial analysis interests, Didier affectionates their use of interoperability settings with conflation models for cross-scales integrated applications using web services.

Dr. Konstantinos Daras is a human geographer with expertise in Quantitative Methods and Geographical Information Systems. His research interests include the way in which internal migration varies between countries around the world and the use of advanced techniques (Graph theory Visualisations, Geometric Algorithms and GPU programming) for developing innovative methods of regional aggregation.