Statistical issues in the assessment of urban sprawl indices [1]

Daniela Cocchi, Linda Altieri
Statistical Sciences Dept., University of Bologna, daniela.cocchi@unibo.it; linda.altieri@studio.unibo.it

Marian Scott, Massimo Ventrucci
School of Mathematics and Statistics, University of Glasgow, marian.scott@glasgow.ac.uk, massimo.ventrucci@glasgow.ac.uk

Giovanna Pezzi
Experimental Evolutionistic Biology Dept., University of Bologna, giovanna.pezzi2@unibo.it

Abstract: Urban sprawl is a hotly debated issue, even if a universally agreed definition does not exist. Its evaluation on spatial data is very important, but the properties of commonly used landscape and sprawl indices have to be assessed, and their performance on raster maps at different pixel resolutions checked, in order to better understand the uncertainty and reliability of results.

Keywords: urban sprawl indices, land cover, pixel resolution, raster aggregation rules, spatial dependence indices.

1. Introduction

Urban sprawl is an important issue for biologists, urban specialists, planners and statisticians, and also for official statistics, both in developed and new developing countries. A universally accepted, well established definition of urban sprawl does not exist, but one of its fundamental properties is to capture uncontrolled and inefficient urban dispersion, accompanied by low building density. Urban sprawl usually occurs when urban planning is not well managed; among its consequences are high average transport costs, soil sealing, pollution (Bhatta *et al*., 2010). Three main types of urban sprawl are currently under study: the monocentric form (one core city surrounded by sprawled suburbs), the polycentric form (more than one core city) and the decentralised pattern (no city centre).
Various measurement methods have been proposed in recent years (see a review in Bhatta *et al*., 2010); some of them are absolute (based on the choice of a sprawl threshold for a selected index), other relative (comparison-based). A very popular sprawl index is Shannon's entropy, but the literature advises that a set of complementary indices to integrate information is created to give a more precise idea of this complex phenomenon. Each index is calculated with reference to a certain spatial extent and a certain spatial data resolution, and measures can be compared over space/time.

[1] Work supported by the project PRIN 2008: New developments in sampling theory and practice, Project number 2008CEFF37, Sector: Economics and Statistics, awarded by the Italian Government.

Statistics can address the sprawl issue in many ways, especially by evaluating the most common recent sprawl indices, assessing their properties, uncertainty and behaviour on raster datasets. Our aim is to identify a suitable set of sprawl indices with good properties and the ability to distinguish among the three sprawl forms; additional information comes from the study of indices at different aggregation levels, following the two most commonly used aggregation methods: the majority and the random rule (He *et al*., 2002).
In our study we have used official EEA land cover data, from the CORINE Land Cover programme (http://eea.europa.eu). They are collected from nearly all EU countries and consist of vector data; the data are then rasterised to 100x100 and 250x250m pixel resolution; a binary raster dataset is also derived, which divides the land into urbanised and non-urbanised zones.

2. Motivation of simulation and empirical studies

Starting from the same elementary data, indices of urban sprawl can assume different values according to the level (*i.e.* pixel dimension) and the aggregation method. We started with a simulation study, necessary for assessing the non linearity of the problem under study, then we used the real dataset mentioned above to detect sprawl occurrence. Both studies were run on raster binary data.
We chose a small set of spatial and landscape indices (*i.e.* we do not exploit information on population, transport, pollution …) and assessed, by simulation, their statistical properties. Each index has a different function: they indicate the existence of sprawl (Shannon's Entropy and Contagion's Index, the last being a measure of clustering-dispersion), the proportion of the territory involved (Simpson's Evenness) and the kind of sprawl (Moran's I, a measure of spatial dependence, because we believe we should find hardly any spatial correlation among pixels in sprawled areas); the interesting ability of Moran's I to identify the type of sprawl has been hypothesized by Tsai (2005) and verified and confirmed by our simulation study. Shannon's Entropy, is defined as $H = -\sum_{i=1}^{S} p_i \ln(p_i)$, where p_i is the proportion of pixels of class i and S is the total number of classes (2 in the case of binary data). It varies between 0 (no sprawl) and $ln(S)$ (maximum sprawl); the usual threshold for sprawl is $ln(S)/2$ (Bhatta *et al*., 2010).
The proportion error has been computed to check the reliability of pixel aggregation in terms of similarity to the original image. We have aggregated both simulated and real datasets to three levels following both rules, to compare the two methods' performance and see how much error in our indices' results they cause.
Our simulation study has reproduced the three sprawl types in various scenarios, generated by an underlying autologistic model (following Hughes *et al.*, 2010) plus Gibbs sampling method. The classic autologistic model is defined as

$$P(Z_i = 1 \mid Z_{-i}, \theta) = p_i = \frac{\exp\left\{ X_i \beta + \eta \sum_{j \in N(i)} Z_j^* \right\}}{1 + \exp\left\{ X_i \beta + \eta \sum_{j \in N(i)} Z_j^* \right\}}$$

where Z_i is the i-th pixel's response, Z_{-i} are all other responses in the grid and the vector θ includes the spatial and attraction parameters; the covariates X_i are the spatial coordinates, weighted through the spatial β parameter, and the autocovariates Z_j are the neighbours' values, weighted through the attraction parameter η. According to this model, the probability of finding urbanisation in the i-th cell depends only on its neighbours' responses (the relationship is controlled by η) and by the pixel location (through the value of β). To create the core city area, we fixed high values for both parameters, while sprawled areas had negative values for η. The neighbourhood extent $N(i)$ has to be fixed in advance, and we chose the 4 nearest neighbours system (Bivand *et al.*, 2008).
In our simulations, we firstly varied spatial and attraction parameters in the model, and, as an alternative, we imposed a kernel structure from the core city area to the periphery, *i.e.* the pixels' responses and the proportion of urbanised cells depend negatively on the Euclidean distance from the city centre; this second, more realistic hypothesis has led to better and more coherent results. For each scenario (9 as a whole), we produced 1000 replications and aggregated them to two coarser levels with both rules (54 scenarios in total). We have then computed the above indices on all replications and resolutions. The same computations have been done on both Emilia Romagna and the city of Bologna (Figure 1) areas (selected from CORINE data) where the original datasets have been aggregated to 500x500,1000x1000 and 5000x5000m pixel sizes.

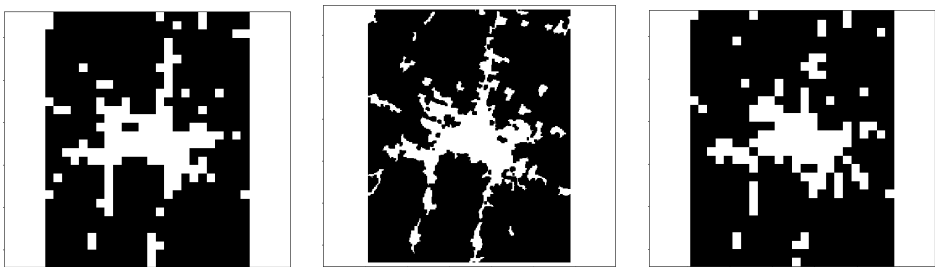

Figure 1. Bologna datasets in the original resolution (100x100m, central panel) and aggregated to 1x1km with the majority method (left panel) and random method (right panel) .

3. Results and comments

Results evaluate indices' stability along aggregation levels and methods, to respectively assess the bias induced by a loss of pixel resolution, and/or using a different aggregation method. The majority rule is a deterministic aggregation method, while the random rule basically draws a simple random sample for each aggregation, starting from the finer resolution data. It appears to be very reliable in the dichotomous case, because the probability of an aggregated pixel falling into one binary class is proportional to the percentage of original pixels in the population of finer elements. The majority method tends to cluster and over-represents the pixels with higher frequency: it is not suitable for detecting dispersion in the data, because it will tend to underestimate it. This has been noted, *e.g.*, in the simulation results for Shannon's Index: after two aggregation steps with the majority rule, the Index did not show occurrence of sprawl, completely contradicting the results from the original data. In conclusion the random aggregation rule is good for measuring sprawl, and leads, in general, to very stable results, *i.e.* more similar to the original, even if its variability always has to be considered.

The variability in indices' measures (in simulation analysis, measured with standard errors and ranges) is higher the coarser the resolution, irrespective of the aggregation method: this suggests it is better to work on the finest resolution possible, even if results are stable over aggregation. The proportion error, which is a classification error, also increases when the resolution becomes coarser, but this tendency is stronger with the majority rule than the random rule. Simpson and Shannon's measures lead to analogous results because they are both based on urbanised pixels' proportions; since no information on pixels' spatial distribution is used, we suspect that they are not the best in identifying sprawl. They are stable when aggregating with the random rule, and, with real data, they identify sprawl in Bologna but not in Emilia Romagna, which suggests that these indices are not reliable on such a wide spatial extent: sprawl is a metropolitan, not a regional, problem. The contagion measure, which is a modified entropy measure containing some information on pixels' neighbourhood, is consistent with Shannon's I, remains stable with the random rule and states that there is sprawl in Bologna. Moran's I is the only index which is able to distinguish (in our kernel simulation study) among the three sprawl types, as shown in Figure 2; on real data it detects occurrence of monocentric sprawl in Bologna, as supported by its map visualization (Figure 1).

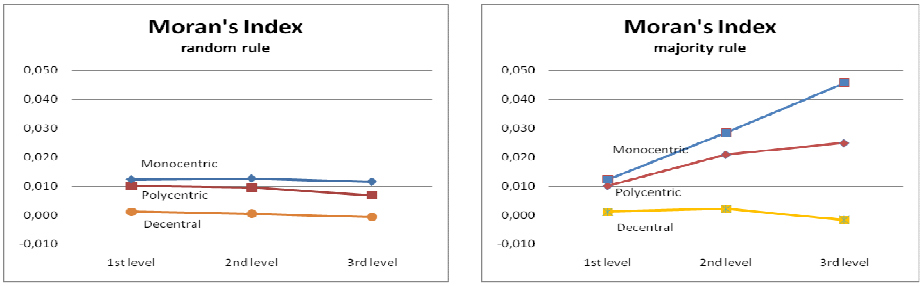

Figure 2. Kernel simulation study; Moran's I at various scenarios and aggregation levels, with both aggregation rules.

In conclusion, the chosen set of indices is suitable for measuring urban sprawl and for identifying the type of dispersion; a further step will be the construction of a unique, composite indicator to identify and quantify such spatial sprawl. As CORINE original data are in vector form, indices such as Simpson, Shannon and Moran's (for binary data) Index should be also computed on vector data to check consistency among results.

References

Bhatta, B., Saraswati, S. and Bandyopadhyay, D. (2010). Urban sprawl measurement from remote sensing data. *Applied Geography*, 30, 731–740.

Bivand, R.B., Pebesma, E.J. and Gómez-Rubio, V. (2008). Applied spatial data analysis with R, UseR! Series, Springer.

He, H.S., Ventura, S.J. and Mladenoff, D.J. (2002). Effects of spatial aggregation approaches on classified satellite imagery. *International Journal of Geographical Information Science*, 16(1), 93-109.

Hughes, J., Haran, M. and Caragea, P. (2010). Autologistic models for binary data on a lattice. Environmetrics.

Tsai, Y.H. (2005). Quantifying Urban Form: Compactness versus 'Sprawl'. *Urban Studies*, 42(1), 141–161.