Socio-economic cluster detection with Spatial Scan Statistics. Case study: services at intra-urban scale

Devis Tuia[1], Christian Kaiser[2], A. DaCunha[2], M. Kanevski[1]

[1]Institute of Geomatics and Analysis of Risk, University of Lausanne, Switzerland
Telephone:+4121 692 35 38
Email: devis.tuia@unil.ch

[2]Institute of Geography, University of Lausanne, Switzerland
Telephone: +4121 692 30 99
Email: christian.kaiser@unil.ch

1. Introduction

In history, planning has been led by several philosophies, aiming on one hand at the production of self-sufficient small scales urban nuclei (Giovannoni, 1931) and on the other hand at the functioning of the city as an entity, i.e. favouring networks planning and hierarchical development. The challenge of sustainable development obliges these concepts to be reviewed and forces the planners to integrate the multi-scale nature of urban space in the decision process.

In that sense, detection of clusters of urban activities is a crucial question in urban planning: the emergence of specialized areas within a city is an indicator of urbanization process and can also be useful for economical analysis. The lack of a certain class of activities can produce supplementary transport flows and/or emissions and influence the quality of life. Several works have been led in order to characterize socio-economic indices and urban supplies networks behaviour, the more often using scaling/Zipf laws (for instance Kühner et al. 2007), but the spatial distribution of activity clusters is rarely considered.

In this contribution, a recently developed cluster detection model, the Spatial Scan Statistics (SSS), is used to detect over- and under densities of a class of services: the hotel business, including hotels and restaurants. This method has the advantage of being independent from the hypothesis of spatio-temporal stationarity.

2. Spatial scan statistics

The spatio-temporal scan statistics have been introduced in the field of public health in the purpose of detecting clustering and emergence of epidemics (Kulldorf 1997, Kulldorf et al. 1998, Kulldorf et al. 2006). The method analyzes events considering them to belong to a random Poisson distributed point process.

2.1 Spatial Poisson Scan Statistics

In the Poisson model, a circular moving window scans the area under study. Each sub-area is called a zone z_i. Each zone owns a number of events c_i conditioned on a certain population p_i, given by the sum of the values of the spatial entities (for instance, municipalities centroïds or hospitals coordinates) contained in the scanning window. Therefore, SSS works on local neighbourhood logic.

When events and population have been computed for a zone, a likelihood function (LF) under the Poisson hypothesis is computed (see Kulldorf (1997)), comparing the events within (subscripts *i*) and outside the zone (subscripts *k*):

$$LF_i = \left(\frac{\left(\frac{c_i}{p_i}\right)^{c_i} \left(\frac{c_k}{p_k}\right)^{c_k}}{\left(\frac{c_{tot}}{p_{tot}}\right)^{c_{tot}}} \right) I \tag{1}$$

Where *I* is an indicator function

$$I \begin{cases} 1 & if \quad \frac{c_i}{p_i} > \frac{c_k}{p_k} \\ 0 & otherwise \end{cases} \tag{2}$$

The indicator function discards results when the ratio counts/population is bigger outside of the zone than inside.

The analysis is performed for every zone within the region, the window taking every spatial entity as a centre and considering different radius (Figure 1). Thus, the local windows principle avoids sensibility to nonstationarity/trends and allows the algorithm to learn directly from the data the size of clusters that must not be specified in advance (Coulston and Riiters 2003, Tuia et al. 2007).

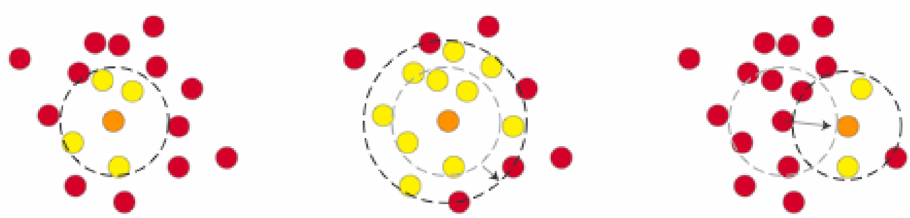

Figure 1. Scanning window principle. (Left) zone z_1, (centre) zone z_2 by increase of the radius, (right) zone z_3 by change of central spatial entity.

The aim of the method is to find the zones maximizing the LF, i.e. the zones where the departure from spatial randomness is maximal. These zones are called the most likely clusters (MLC).

Since the MLC have been highlighted, their significance has to be tested. The procedure implies Monte Carlo simulation: a certain number of data simulations (usually 1000) respecting the null hypothesis H_0 of spatial randomness is generated and the LF of each replication is compared to the most likely clusters of the real data. If real clusters are included in the 5% of the replications, H_0 hypothesis is rejected and the cluster found are significant.

The spatiotemporal SSS follows the same procedure, but, instead of considering spatial circles, spatiotemporal cylinders analyzing different spatial and temporal neighbourhoods are used.

3. Data

The data used come from the Swiss Census, and more specifically from Population and Firms Census. Both datasets have been extracted and aggregated following a grid with a 200 meters resolution: this grid represents the spatial entities. As temporal extent, the analysis has been limited to the emergence of clusters during the 1990-2000 period.

The area under study chosen is the city of Lausanne (Western Switzerland) and its agglomeration (Figure 2): this agglomeration, counting 300’000 inhabitants for 68 municipalities. The performing transport network guarantees fast and reliable accessibility to the centres of the agglomeration.

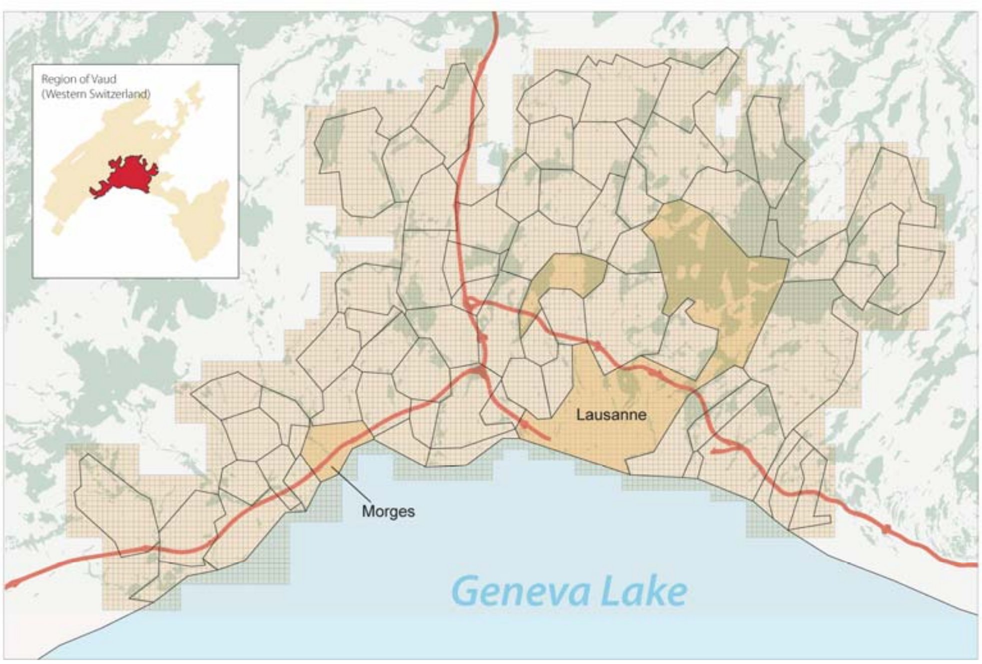

Figure 2. Agglomeration of Lausanne. Brown cells constitute the spatial entities of the study. (Source: Swisstopo)

4. Discussion

First, SSS has been applied taking into account the number of firms in the sector (weighted by the size of the firms) as events and the total population as population. Figure 3 shows the spatial repartition of the emerging clusters: red colour scale shows the high-rate clusters, while blue colour scale shows the low-rate clusters.

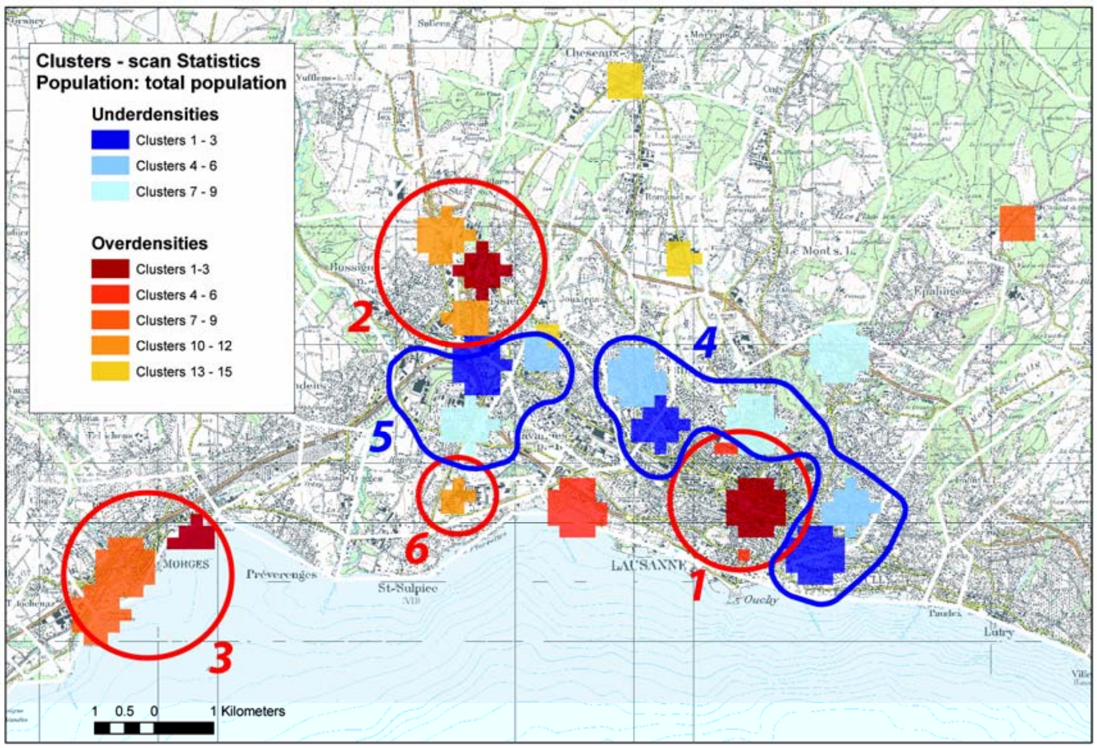

Figure 3. SSS (only clusters with significance level of 0.001 are represented).

Clusters 1 and 3 highlight the Lausanne and Morges city centres, hosting several hotels and restaurants. Cluster 2 shows the industrial region of Crissier, where commercial centres increase the number of services in an area characterized by low population. Residential areas in interaction with clusters of high rates, characterize clusters 4 and 5. Cluster 6 is a particularity of the region: it corresponds to the Institute of Technology of Lausanne, which do not have any residential population, but is characterized by several services related to students.

The use of resident population alone as conditional information introduces a bias to the analysis. For a variable such as the number of hotels and restaurants, the tourism is an essential component that is not taken into account. Cluster 2 of high rates shows well this problem: this zone, characterized by small population, is supposed to have a low number of events; nonetheless, because the zone is a tourist area, the number of services is 1200 times greater than expected! In order to solve this bias, the population parameter has been replaced by the sum of the resident population and of the number of employments in the tertiary sector. Doing this, the conditioning variable takes into account the employment. The results of such correction are shown in Figure 4.

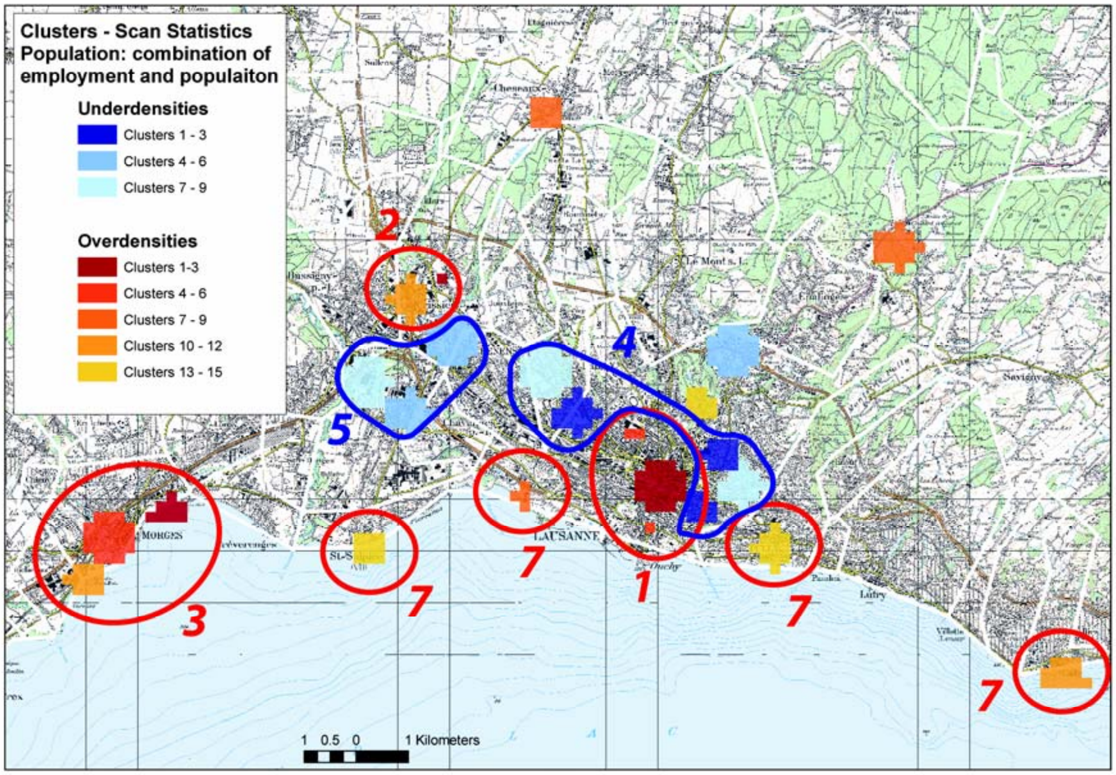

Figure 4. SSS using the sum of population and employment (only clusters with significance level of 0.001 are represented).

The use of the hybrid population has an important impact on the clusters highlighted by the previous analysis: Clusters 1, 2 and 3 are confirmed, but they have decreased in importance (the employment has weighted the LF). Cluster 4 grows in importance: the presence of important employment centres such as the regional hospital make the peripheral low-rates clusters become more important. Cluster 5 loses in importance, because the population of the area has grown less than expected by adding the employment data. Cluster 6 simply disappeared. Tourist areas (Cluster 7) in front of the lake appeared.

5. Conclusion

Spatial Scan Statistics is a powerful method to detect emerging clusters in space and time. Such a method could be used in urban studies and planning to detect areas where a lack of services could lead to forced trips or to a loss in the quality of life. In this contribution, SSS have been applied to the question of hotels and restaurants in the agglomeration of Lausanne and allowed to detect and interpret clustering. The inclusion of employment-related variables in the conditioning parameters led to a better modelling of the situation of the last 10 years in the region. This first application is very promising and the method shows good performance for the specific problem. Ongoing studies focus on update of the conditioning information with expert knowledge.

6. Acknowledgements

This work has been partly supported by the Swiss National Foundation (project “Urbanization Regime and Environmental Impact: Analysis and Modelling of Urban Patterns, Clustering and Metamorphoses”, n.100012-113506).

7. References

Coulston, J.W, Riiters, K.H, 2003, Geographic analysis of forest health indicators using spatial scan statistics, *Environmental Management*, 31: 764-773.

Giovannoni, G, 1931, Il miglioramento dei nuclei interni. La teoria del diradamento. In: Giovannoni, G, *Vecchie città e edilizia nuova*, Roma, 248-280.

Kühnert, C., Helbling, D, West, G.B, 2007, Scaling laws in urban supply networks, *Physica A*, 363: 96-103.

Kulldorf, M, 1997, A spatial scan statistic. *Communications in Statistics*, 26: 1481-1496.

Kulldorf, M, Athas, W, Feuer, E, Miller, B and Key, C, 1998, Evaluating clusters alarms: A space-time scan statistic and brain cancer in Los Alamos. *American Journal of Public Health*, 88, 1377-1380.

Kulldorf, M, Song, C, Gregorio, D, Samciuk, H, DeChello, L, 2006, Cancer maps patterns: are they random or not? *American Journal of Preventive medecine*, 30, 37-49.

Tuia, D., Ratle, F., Lasaponara, R., Telesca, L., Kanevski, M., 2007, Scan Statistics analysis for forest fires clusters, *Communications In Nonlinear Sciences and Numerical Simulations*, in press.