Available online at www.sciencedirect.com

ScienceDirect

Procedia Environmental Sciences 27 (2015) 66 – 69

Procedia Environmental Sciences

Spatial Statistics 2015: Emerging Patterns - Part 2

R as a GIS: illustrating scale and aggregation problems with forest fire data

Romain Louvet[a]*, Jagannath Aryal[b], Didier Josselin[a,c], Cyrille Genre-Grandpierre[a]

[a]*UMR ESPACE 7300, 74 rue Louis Pasteur, 84029 Avignon Cedex, France*
[b]*University of Tasmania, School of Land and Food, Private Bag 76, Hobart, Tasmania, 7001, Australia*
[c]*Laboratoire d'Informatique d'Avignon, LIA, 339 chemin des Meinajariès, Agroparc BP 91228, 84911 Avignon cedex 9, France*

Abstract

Using the R language as a GIS applied on forest fire data in South of France, the goal of the research is to emphasize how spatial statistics may depend on the areal units chosen. First, we propose to map the forest fire data at different scale levels based on administrative boundaries. Second, we measure the MAUP by showing scale sensitivity in descriptive statistics and in regression analyses. Finally, although many tools can be used for vector or raster data aggregation and mapping, we discuss why we choose R as a primary analysis tool and R added-value.

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns committee

Keywords: Modifiable Areal Unit Problem, forest fire, R

1. Introduction

"[T]here are no rules for areal aggregation" [1]. Although this is no longer entirely true, many studies still use administrative areal units at only one scale [2], ignoring scale and aggregation problems. Known as the MAUP (Modifiable Areal Unit Problem), this issue is one of the main current challenges in spatial data analysis. This aggregation problem is considered to occur when the way that data are arranged into areal units has a significant impact on observation. Areal units' data arrangement has two facets, corresponding to two effects of the MAUP:

* Corresponding author. Tel.: +33-6-77-43-64-19
E-mail address: romain.louvet@alumni.univ-avignon.fr.

Peer-review under responsibility of Spatial Statistics 2015: Emerging Patterns committee
doi:10.1016/j.proenv.2015.07.118

scale and delineation [1]. The scale effect is defined as a change in a result obtained from the same data when they are aggregated at different scales and when areal unit sizes increase. The delineation or boundary effect is due to a change in a result from the same data when they are aggregated at a same scale with different areal unit boundaries and while the number of areal units is constant. In other words, the MAUP exists when spatial data are differently aggregated, therefore raising questions about certainty of spatial statistics [3].

Since its discovery attributed to Gehlke and Biehl [4] this problem has been observed in seminal works written by Robinson [5] and Openshaw and Taylor [6]. Even if it was described many times only a few practical solutions are implemented, and no general solution is agreed upon [3,7]. The simplest solution would be to use only disaggregated data. Unfortunately, these data are rarely available and are often impaired due to their low power of communication (e.g. it is better to draw maps with a well-known set of areal units). This leads to a certain temptation to ignore the spatial dimension when dealing with data points, or getting reliable rates for policy formulation and implementation [7,8]. Other solutions have been proposed with different approaches. A first solution consists in adapting statistics index formula, for instance a correlation weighting based on areal unit sizes [9]. Considering the MAUP as a "tool" rather than a problem can also be a way, since it is closely related to spatial structure of variance [7]. Then we should use it to find a relevant set of areal units based on a resulting optimization [1]. One way to do so is to use autocorrelation or geographical weighted regression (GWR) in order to generate a set of areal units that would show more information over space [10]. But this optimization approach seems to contradict basic objective scientific methodology. Another way is to use a set of grouping variables known at an individual level, and then adjust a variance-covariance matrix of aggregated data in order to select the set of areal units that shows the closer similarity with underlying individual level of variance [11,12,13]. Sensitivity analysis can comprehend the MAUP by comparing the results of areal unit sets to known individual level statistics [8]. This last solution has one major flaw: it needs individual level data. Based on Bayesian statistics, a solution to this problem could be to measure the sensitivity to the MAUP by using a Bayesian estimator of simulated random individual level data [14]. A similar approach is based on spatial resampling to eliminate the support effect [15]. These solutions are based on the assumption that the MAUP only affects nonrandom data [1].

Another sensitivity measure deals with the MAUP by comparing different sets of areal units altogether, in order to select only results which are similar at many scales or with different boundary sets. This kind of results should induce less uncertainty and have more meaning because their consistency makes them resistant to the way that data are aggregated [3]. This paper proposes to implement this last solution as an exploratory tool programmed in R language and to test it on forest fire data and their driving factors.

2. Methods and data description

Our objective is to build a method that can be used independently from the topic or the territory of the study. We choose to test it on forest fire causes in order to compare our methodology and our results with a previous study [2]. South of France, as other Mediterranean regions in Europe, is regularly and strongly affected by forest fire [2]. Forest fire process is a complex matter, mainly because it is a "natural" disaster driven by human factors. Ganteaume and Jappiot's study [2] showed interesting results about this fact and their data are available. Moreover, they choose to work at the French *départements* scale with data available at a more disaggregated level. Choosing the French *départements* scale is not that arbitrary since wildfire suppression policies are supposed to be conducted at this scale. However, using seemingly arbitrarily determined, therefore modifiable, areal units is an important methodological problem, independent from administrative unit related considerations.

Based on French political territorial units, our primary areal unit set was a shape file of 3569 *communes* (LAU2), which was then aggregated into 475 *cantons* (LAU1), 245 *EPCI* (public organization of intercommunal cooperation), 43 *arrondissements*, 15 *départements* (NUTS3) and 4 *régions* (NUTS1). We only studied the scale effect since we had only hierarchical and non-overlapping areal unit sets, and no other types of areal units such as grid frameworks or Voronoï tessellations. These aggregation codes come from *Géofla* database 2014 (IGN), except the EPCI 2014 coming from *collectivites-locales.gouv.fr*. We have a total of 4 dependent variables: forest fire number of occurrences and surfaces during the 1997-2013 period from the *Prométhée* database. These data were "normalized" (we tried to align them to a Gaussian distribution) as occurrence densities (fire occurrences divided by area) and fire surface rates (total fire surfaces divided by total area). Number of fire occurrences and total of fire

surfaces were obtained prior to the R program using Excel to clean the raw dataset. The dependent variables were log-transformed in order to follow a parametric pattern [2]. We used 4 explanatory variables: population density (*Géofla*, IGN, 2014), road density based on IGN's *Routes 500* 2012 dataset (roads length divided by area), forests land cover density and shrub and/or herbaceous vegetation associations land cover density (specific land cover divided by total area) from Corine Land Cover 2006 statistics. ArcGIS was used to get road density by intersecting road arc with *communes'* polygons and then sum road arc length with a spatial join to the communes.

Data processing was implemented using R programming and preferentially functions in order to be applied on different data sets. This processing includes: loading specific packages to handle spatial data (i.e. *rgeos* and *rgdal*), loading the data (shape file, variables) at the most disaggregated level, aggregating selected variables according to a list of grouping variables, processing normalized variables at each scale separately, creating a new shape file for each scale level, returning a list of spatial objects corresponding to scale levels. Then, based on this list of spatial objects, a script produces maps, descriptive statistics and plots.

3. Results and discussion

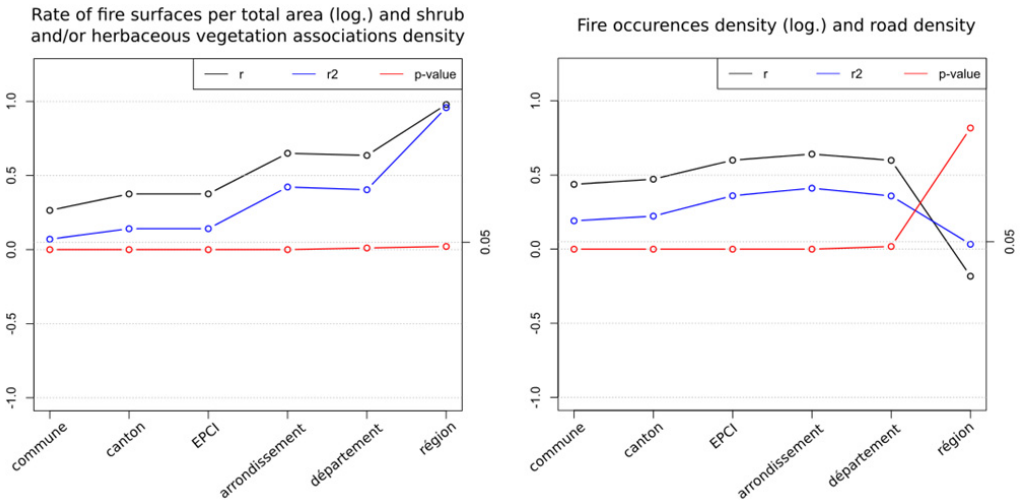

Fig. 1. r, r^2 and *p-value* variations at different scale levels

The primary result of this methodology is to automatically and quickly produce a total of 276 diagrams, 12 summary tables, 48 maps, and six shape files, based on a single shape file, a dataset of eight variables, and political geographical units' codes. Diagrams include usual histograms, box-plots and scatter-plots, but also diagrams representing how different statistical indices or results vary through scales (Fig. 1). Therefore, this provides a tool for exploratory analysis which enables to monitor possible effects of the MAUP on the results.

Mapping systematically every variables at six scales shows how the MAUP affects the data by erasing local variance (we use quantile classification for its convenience when comparing maps). We also observed the scale effect on variance with a variation between minimum and maximum values which can be more than 90,000 times with our data. Excluding extremely big variations, the total mean of variance variation is still 14.8. We observed a difference between "normalized" and "non-normalized" variables: non-normalized variables seem to be more sensitive to the scale effect. The mean variation of the mean is 6.1 for normalized data and 892.3 for non-normalized data. We obtained similar results with r^2 coefficient variation: mean of 3.8 for normalized data and 8 for non-normalized data (we used only r^2 with a *p-value* <= 0.05). This difference because of normalization is not very surprising since the scale effect is directly linked to the size effect. But so important variations, more or as much important of the mean vs. r^2 coefficient, are surprising since the MAUP effect on average and variance is supposed to be less severe than for correlation coefficients [7]. Moreover the variance of non-normalized data increases with larger areal units unlike normalized data. Since it is recommended to use less aggregated data because they are

supposed to be closer to a maximum of variance, these observations could be strong assumptions for recommending using only normalized variables in order to limit the MAUP.

The secondary result could be the capacity of the process to produce more robust correlation analysis. By showing statistical variation through scales, we can detect tendencies, anomalies, and select information based on their consistency at different scales. Indeed, results showed that some relationships can be relatively stable, while others appear to be very sensitive to the scale effect [16]. Also, it is generally assumed that we observe increasing r with increasing areal unit sizes [17]. We obtained the same results: a main tendency to an increasing r^2 with increasing areal unit sizes, some relations were stable, some consistent in increasing process (Fig. 1) and others very erratic. For instance, correlation between fire occurrence density and road density seems to be sound (Fig. 1) since it is relatively consistent at every scale level. The variation between maximum and minimum r^2 with a *p-value* <= 0.05 is 2. On the contrary, the relation between fire surface densities and shrub vegetation densities is problematic, since the maximum significant r^2 is equal to 14 times the minimum significant r^2. These results are similar to the results from Ganteaume and Jappiot's study [2].

This approach to manage the MAUP using R as a GIS shows promising methodological results which could be deepened by further works. Although many tools can be used for vector or raster data aggregation and mapping (e.g. ArcGIS, Qgis, Grass, GeoDMA, eCognition), R is a powerful tool which has numerous advantages, especially if considering the nature of the problem we posed in this paper. R is a free and versatile software/programming language, becoming the *lingua franca* for data analysis. It can be used for integration of GIS, statistics, plots, and automatic mapping into a single workflow, and sharing scripts along with analyses [18]. Considering the MAUP, R includes robust packages to handle spatial data and may be even superior to GIS software in aggregation and plotting sequentially, including maps (for instance, let us note the package *RgoogleMaps*). Unfortunately R is not so good for interactive use and therefore limited for exploratory analyses. As a next step, the approach adopted in this research could be completed by the study of the boundary effects and by implementing proposed solutions such as guided choices of an optimized areal unit sets (e.g. based on variance-covariance matrix and on Bayesian statistics).

References

[1] Openshaw, S. (1984). 'The modifiable areal unit problem', Norwich: Geo Books, CATMOG, 38.
[2] Ganteaume, A., Jappiot, M. (2013) 'What causes large fires in Southern France', Forest Ecology and Management, Elsevier, 2013, 23 p.
[3] Fotheringham, A. S., Brunsdon, C., Charlton, M. (2000) *Quantitative Geography: Perspectives on Spatial Data Analysis*, SAGE
[4] Gehlke, C.E., Biehl, H. (1934) 'Certain effects of Grouping upon the Size of the Correlation Coefficient in Census Tract Material', Journal of the American Statistical Association, Supplement 29: 169-70.
[5] Robinson, W.S. (1950) 'Ecological correlations and the behaviour of individuals', American Sociological Review, 15, 351–7.
[6] Openshaw, S., Taylor, P.J. (1979) 'A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem', in N. Wrigley (ed.), Statistical Applications in the Spatial Sciences. London: Pion. pp. 127-44.
[7] Swift, A., Liu, L., Uber, J. (2008) 'Reducing MAUP bias of correlation statistics between water quality and GI illness', Computers, Environment and Urban Systems 32, n°2. 134-48.
[8] Arsenault, J., Michel, P., Berke, O., Ravel, A., Gosselin, P. (2013) 'How to choose geographical units in ecological studies: Proposal and application to campylobacteriosis', Spatial and Spatio-temporal Epidemiology 7. 11-24
[9] Robinson, A.H. (1956) 'The Necessity of Weighting Values in Correlation Analysis of Areal Data', Annals of the Association of American Geographers, 46: 233-6.
[10] King, G. (1997) *A solution to the ecological inference problem. Reconstructing individual behaviour from aggregate data*, Princeton University Press.
[11] Steel, D.G., Holt, D. (1996) 'Rules for Random Aggregation', Environment and Planning A, 28: 957-78.
[12] Holt, D., Steel, D.G., Tranmer, M., Wrigley, N. (1996) 'Aggregation and Ecological Effects in Geographically Based Data', Geographical Analysis, 28: 244-61.
[13] Tranmer, M., Steel, D.G. (1998) 'Using Census Data to Investigate the Causes of the Ecological Fallacy', Environment and Planning A, 30: 817-31.
[14] Hui, C. (2009) 'A Bayesian Solution to the Modifiable Areal Unit Problem', in Foundations of Computational Intelligence Vol. 2. Editors: A-E. Hassanien, A. Abraham, F. Herrera, 175-96. Studies in Computational Intelligence 202. Springer Berlin Heidelberg.
[15] Mafhoud, I., Josselin, D., Fady, B. (2007) 'Effect of the aggregation process on the diversity assessment toward the "pertinent scale"'. AGILE 2007, conference proceedings.
[16] Fotheringham, A.S. Wong, D. (1991) 'The Modifiable Areal Unit Problem in multivariate Statistical Analysis', Environment and Planning A, 23: 1025-44.
[17] Blalock (1964*) Causal Inferences on Nonexperimental Research*, Chapel Hill, NC: University of North Carolina Press.
[18] Commenges, H., Beauguitte, L., Buard, E., Cura, R., Le Néchet, F., Le Texier, M., Mathian, H., Rey, S. (2014) *R et espace, Traitement de l'Information Géographique*, Groupe ElementR, Framabook, Paris.