Effect of the aggregation process on the diversity assessment: toward the 'pertinent scale'

Ilene Mahfoud * , Didier Josselin* and Bruno Fady**
*UMR 6012 ESPACE CNRS – Université d'Avignon
74, rue Louis Pasteur 84029 Avignon cedex 1, France
didier.josselin@univ-avignon.fr, mahfoud_ilene@yahoo.fr
**INRA d'Avignon, Unité de recherche forestière méditerranéenne
Site Agroparc Domaine Saint Paul 84914 AVIGNON cedex 9, France

INTRODUCTION

Lots of the studies about landscape ecology require currently aggregated data to analyse the phenomena in space and time. This evolution led the scientists to work on larger areas (*i.e.* lower geographical scales), such as ecological regions, to assess the landscape dynamics, biodiversity, or more generally, the global change (Jelinski & Wu, 1996).

The Modifiable Areal Unit Problem (MAUP) has been observed and studied for the Thirties, more recently in the field of landscape ecology. This problem points out the difficulty to provide a synthetic and reliable statistical index according to a given scale, because the measure will be more or less sensitive to the source of spatial information itself. It involves indeed problems of accuracy, boundaries and levels of aggregation. Practically, the MAUP has been enhanced thanks to the development of GIS and remote sensing software. That favors a rather easy handling of the spatial data without necessarily asking the previous question of their real reliability. These different issues have made the MAUP a key problem for many scientists who have produced a large and interesting research about it (Openshaw, 1984, Rastetter *et al.*, 1992 ; Reynolds, 1998).

The MAUP can be divided in several linked aspects. Firstly, at a given scale or aggregation level, the information for each spatial object can be considered as an average of the information encoded in the basic entities composing the object. This infers a (risk of) loss of information depending on the aggregation level (Reynolds, 1998). Secondly, the aggregation method which is used plays a non negligible role in the estimate provided at a given scale (Openshaw, 1984 ; Dusek, 2005). Thus, with a constant number of initial entities, the values of different statistical models (correlation, for instance) can significantly vary. We can add to these two well-known points the effect of the statistical efficiency. Indeed, the number and the spatial and statistical distributions of the data have such an influence on the method robustness that one can wonder how important is the statistical part of the problem, its spatial expression being probably a visible consequence rather than the deep cause of the phenomenon.

Several studies aimed to test the sensitivity of the analysis to the aggregation effect. The first observation has been stated in 1934 by Gehkle and Biehl, who noticed a relation between the correlation estimates and the spatial levels of the data. Similar results have been observed by Yule and Kendall in 1950 concerning the potatoes and wheat yields correlation for 48 English regions. Robinson then demonstrated (1950) that the values of correlation increase when the number of observations decreases or the size of the spatial unit grows. In 1976, Clark and Karen also confirmed these effects.

Other scientists showed that the MAUP has effects in various application contexts. In 1994, Marceau *et al* verified the impact of image spatial resolution and aggregation levels on the accuracy of remote sensing data classification. This work found out that the accuracy estimated per class has been considerably affected by these two factors. So remote sensing data set as a particular case study of the MAUP, where the size of the pixels and the way to aggregate them, within a regular spatial partition, must be handled with a great caution (Jelinski et Wu, 1996 ; Wu, Gao & Tueller, 1997).

Actually, a few ways have been proposed in the literature to take into account aggregated data while trying to avoid the aggregation effect. A first approach consists in setting aside the traditional statistical methods sensitive to the MAUP and to work on more thematical analysis (Jelinski et Wu, 1996). Openshaw (1984) looked for modifying the spatial entities to make them more reliable, more adequate and in a certain way, 'non modifiable'. This solution requires a deep knowledge of the handled geographical objects. A variance analysis can also be useful to provide 'optimal' partitioning by maximizing the variance between the areas while minimizing the one in each entity. This method can fail when applied on different variables and is sometimes not easily reproducible. Other research can be developed to assess the weight of the aggregation effect in the value of the statistical estimates, using notably mathematical simulations (Amrhein, 1995, Reynolds, 1998). That is the way we propose to proceed, by studying the internal diversity of the aggregated data.

METHODOLOGY

Vegetation and diversity index assessment

In our research, we chose to tackle the MAUP by testing the robustness of the Shannon entropy related to the aggregation effect. This index is indeed widely used in spatial analysis. In our study, it is calculated using an image from SPOT 5 (resolution: 5 meters), whose pixels are aggregated at several levels of interwoven scales. The objectives are:

- to avoid, or at least reduce, the impact of the aggregation on the value of the calculated index;
- to look for the 'most pertinent scale' for which the diversity index should be processed and relevant, at least the less sensitive possible to the aggregation effect, or the most discriminant.

The source image is a part of a SPOT 5 panchromatic multispectral image covering the 'Mont du Ventoux', a little mountain of the South-East of France (in the Vaucluse region). A Normalized Difference Vegetation Index (NDVI) has firstly been computed using the software GRASS 5, allowing to make an analysis of the biomass diversity.

The Normalized Difference Vegetation Index is computed as:

$$NDVI = (B3 - B2) / (B3 + B2)$$

where B2 is the value from the red channel sensor and B3 the one from the close infrared.

The Shannon entropy index is often used for describing how equitably is arranged a set of values or individuals. In our case, we apply this index to a classification of biomass. Generally, biologists use it to measure the taxonomic richness in terms of species. The Shannon index (H') is minimal if, for instance, a species is dominant and the other ones are represented by only a few individuals, and maximal when all the species are uniformly distributed. The more equitable the distribution, the higher the index (Frontier, 1983). The formula is:

$$H' = \sum_{i=1}^{m} p_i * \ln\left|p_i\right|$$

with p_i the frequency of the attribute i in the whole set and m the number of possible attributes.

A systematic evaluation of the diversity through the scales

In order to analyse the aggregation effect, we cut the image in several interwoven grids, whose aggregates (of pixels) have an edge of 10, 20, 40, 80, 160 and 320 meters. Then, we computed the diversity index for these 6 partitions. That is to say that we provided the Shannon diversity for all the pixels aggregates, and obtained an image for each scale. These images have different numbers of individuals depending on the aggregation level: this number decreases while the aggregation process runs. We added to these results the mean and the median and the distribution of the diversity values for each whole image. This method allows to assess the effect of aggregation on the diversity index.

The first common example of this process involves the observed data. This provides a first image, corresponding to the vegetation observed diversity (using the NDVI index).

However, the objective to minimize or to avoid the impact of aggregation cannot be reached at this stage. That is why we built a method to identify the part of the diversity which is due to the spatial structure independently from the level of aggregation. Indeed, we needed other images as references corresponding to several cases of typical distributions of the diversity. From the original classified image of the vegetation, we computed 3 images, with the same number of initial pixels and identical map extents.

The second image is completely covered by the same value. In this case, whatever the scale, the diversity is equal to 0 (unique class of value). This case is extremely rare (!) and is interesting for us only to remind what would be a territory without any diversity.

At the apposite, we built for each level of aggregation a third image representing the 'maximum' of diversity. Over the whole image, we drew as many different pixels as possible (there don't exist two identical pixel values indeed). This case corresponds to the maximal diversity and also to a kind of statistical uniformity, because there isn't any dominant value (or species, if we tackle the biodiversity). Other indices might be used to complete the analysis with dominance and variety assessment, for instance (Mahfoud *et al*., 2006).

Finally, we processed for each image a re-sampling of the original pixels of the vegetation index (NDVI) to build a random image under statistical constraints. That is to say the total series of the observed pixels has been re-distributed randomly all over the image. 100 resulting images have been designed and several tests showed that the results of the analysis are similar whatever the randomized image, despite some differences that can exist most of time for low resolutions. We selected one of them to develop our analysis. In this particular re-sampled image, the values obtained are cleared of their spatial structure and autocorrelation, because the pixels have been randomly distributed. In other terms, this image includes the part of the diversity due *(i)* to the size of the image, *(ii)* to the level of aggregation, and also *(iii*) to the set of values of the initial pixels. This fourth image represents indeed an intermediate situation with generally a rather high diversity.

The complementarity of these images (observed, null, 'maximal' and 'randomized' diversities) through the 6 levels of aggregation enables to take into account the MAUP in the diversity assessment at different scales (*cf.* Fig. 1). Let us now order the images by their probable global diversity:

- The image 0 represents the homogeneity and a null diversity (the values are here replaced by the mean of the pixels values of the whole image);
- The image 1 corresponds to the observed values of diversity where the spatial structure remains visible at any scale;
- For the image 2, the same observed pixels as the case 2 are spread randomly all over the area, making a 'random image' under probability constraints;
- The image 3 is the maximal possible heterogeneity (all the pixels are differing from each others).

RESULTS: AGGREGATION EFFECT ON DIVERSITY ASSESSMENT AND 'PERTINENT SCALE'

The analysis of the sensitivity of the diversity measures to the MAUP is firstly broached by evaluating the range and the evolution of the phenomenon through the scales, according to the set of the different images and the associated statistics (mean, median and distribution of the values of the aggregated pixels). Moreover, we compare the cases *1, 2, 3* and try to distinguish the relative parts of the diversity carried by the spatial structure or included in the aggregation process and level themselves. To do that, we cross the mean and the median of the diversity indices calculated using the aggregates of pixels at different scales (*fig. 1*). Then we compute the diversity deviations (*fig. 2*) and ratios (*fig.* 3) between the several images at different aggregation scales and draw the curves in plots. This method enables to extract the aggregation effect and to identify inflexion point(s) corresponding possibly to 'pertinent scales' for assessing the diversity or one of its facet.

The results show that the spatial structure is not the most significant factor to explain the diversity. The level of aggregation, it is not surprising, plays a role, as well as the size of the image and the total number of pixels, that influence the statistics. However its impact is more important than we could expect. Consequently, the diversity estimates must be comprehended with a lot of caution: the weight of the spatial partitioning is decisive in the diversity assessment.

So the Shannon index is very sensitive to the aggregation effect. We observe indeed that the average of the diversity indices increases with the growth of the spatial resolution in all the cases (*fig.* 2). For example, with pixels of 10 meters edges, the diversity values are 1.265 (*case 1*), 1.354 and 1.386 for the *cases 2* and *3* (which is not significantly different at this scale). For a 40 meters resolution, the means of diversity are more than doubled: respectively 3.154, 3.633 and 4.158 (4.155, 4.425 and 8.361 with aggregates of 320 meters edges). The deviation between the *cases 2* and *3* becomes more and more important, while the progression of the deviation between the *cases 1* and *2* remains rather low. If we assume that the difference between *1* and *2* identifies the spatial organization and autocorrelation, this gives an idea of the low part of the spatial structure in the global diversity assessment if we consider the internal diversity of the aggregated data.

Besides, we can notice that, for the *case 1* (observed data), the number of classes in the diversity classification (see the distributions in *figure* 1) increases with the resolution. This behavior seems more marked than for the random *case 2*. In fact, the random process ensures a well-balanced distribution of the pixels on the whole image, that reinforces the probability to generate aggregates with high diversity. We here emphasize the influence of the spatial structure on the number of classes through the scales. This points out the statistical face of the MAUP, due to the influence of the number of classes in a distribution on the entropy assessment.

This effect is also visible when comparing the curves between *cases 1* and *2* (*fig.* 2). The spatial structure provides globally lower values of diversity. The deviation identifies the decrease of the diversity due to a reduction of the spatial autocorrelation. This allows to eliminate the effect of the aggregation level. Let us recall indeed that the *case 2* (randomization of the image *1*) includes only (*i*) the aggregation level, (*ii*) the size and the shape of the image, (*iii*) the values of the observed pixels, although the *case 1* adds the spatial structure (*iv*) to the previous elements (*i, ii* and *iii*).

That leads us to identify what we call a 'pertinent scale', in the sense it corresponds to the spatial resolution for which the deviation between the observation (*case 1*) and its randomization (*case 2*) is maximized. This is the point where the spatial structure has the highest influence compared to the

other factors of diversity: the diversity explained by the space (what we are interested in) may be the most discriminant at this scale. Our assumption is to consider it as the most usable scale to study efficiently the diversity. In our study, this corresponds to a resolution of about 80 meters (edge of the aggregates). This is confirmed by the *figure* 3, where the ratio between these two diversity estimates (ratio *2/1*) is the weakest at this scale.

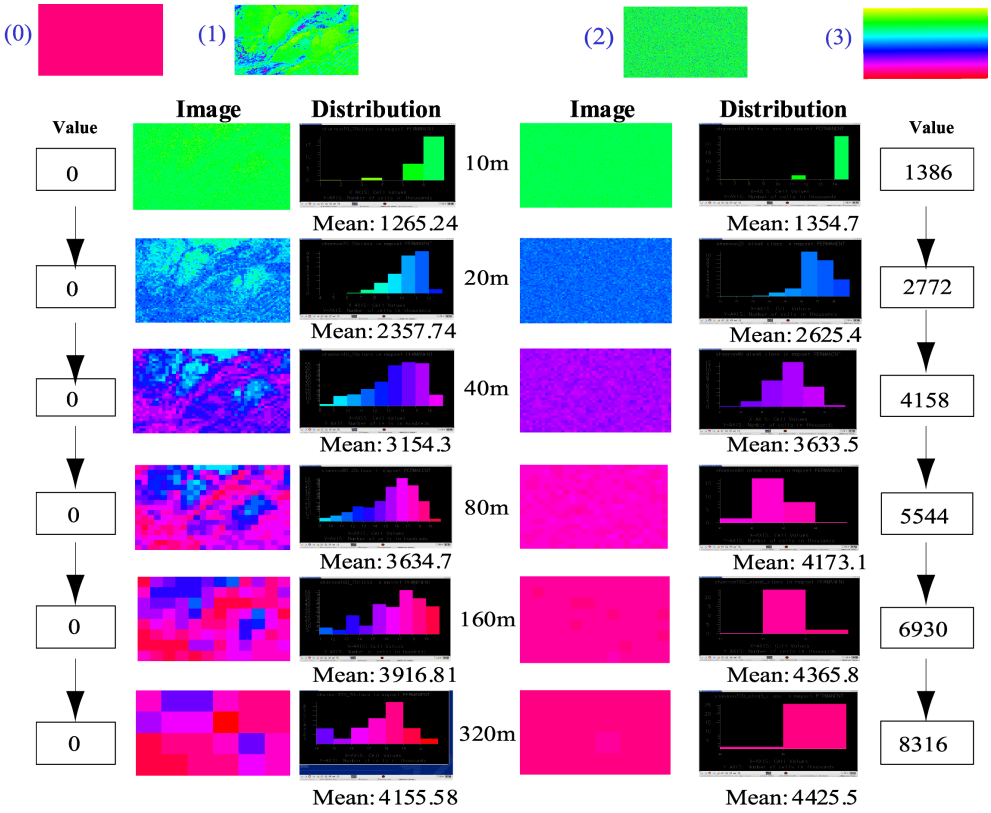

Figure 1: Shannon diversity through the scales

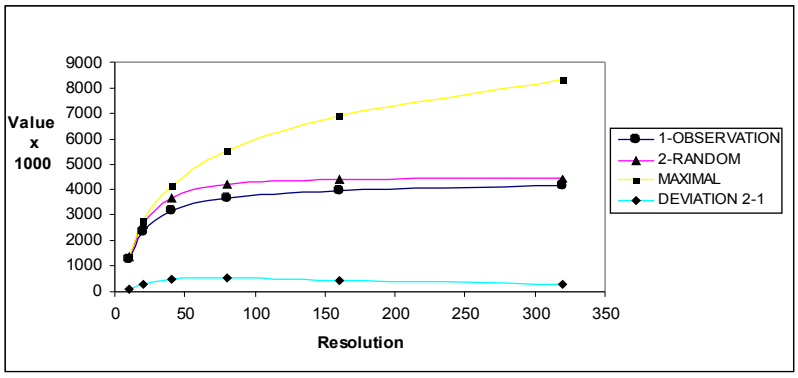

Figure 2: Diversity estimates and their deviations

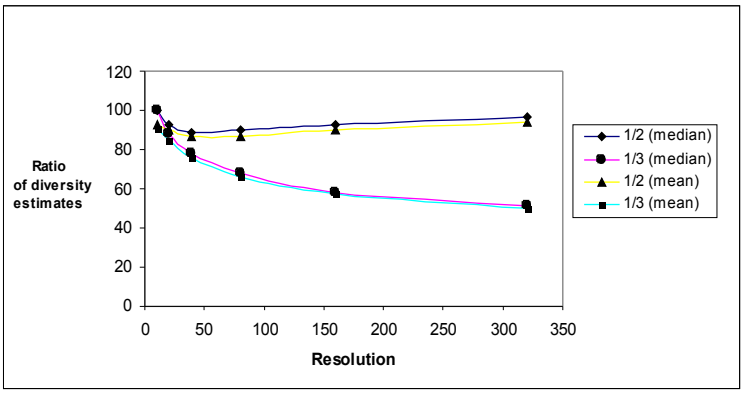

Figure 3: Ratio of diversity estimates

CONCLUSION AND PERSPECTIVES

In this paper we showed that the Shannon diversity is concerned by the MAUP. Then, using a re-sampling, we proposed a method to avoid the aggregation effect. We ended on a way to identify the 'pertinent scale' for which the diversity should be evaluated with a better accuracy. The spatial resolution of 80 meters edges is identified as the 'most pertinent scale' in the particular case of the vegetation diversity applied on the Mont Ventoux (France) using a SPOT 5 image.

Here are our current and further works:

- to compute other indices on the same region to verify if we find the same pertinent scale (Simpson diversity, dominance, richness...);
- to study the relation between the initial resolution and the aggregation effects on the diversity assessment;

- to build a new diversity *ad-hoc* index, more adequate and less sensitive to the aggregation problem;
- to apply this research to the biodiversity evaluation and taxonomic classification using image processing.

BIBLIOGRAPHY

Amrhein, C., « Searching for the elusive aggregation effect: evidence from statistical simulations », *Environment and Planning A*, 27, 1995, p. 105-119.

Baker W., « The r.le Programs, A set of GRASS programs for the quantitative analysis of landscape structure ». 1997, Version 2.2, University of Wyoming, USA. http://grass.itc.it/gdp/terrain/r_le_22.html

Clarrk W.A., Karen L., « The effects of Data Aggregation in Statistical Analysis ». *Geographical Analysis*, 1976, vol. VIII, p.429-438.

Dusek T., « The Modifiable Areal Unit Problem in Regional Economics ». The 45th Congress of the European Regional Science Association, 2005, Amsterdam.

Frontier S., L'échantillonnage de la diversité spécifique. In Stratégie d'échantillonnage en écologie, Frontier et Masson édit, 1983, Paris (Coll. D'Ecologie), XVIII + 494 p.

Gehlke C.E., Biehl, K., « Certain effects of grouping upon the size of the correlation coefficient in census tract material ». 1934, *Journal of the American Statistical Association*, p. 169-170

Jelinski D.E., Wu J., «The modifiable areal unit problem and implications for landscape ecology ». *Landscape Ecology*, 1996, vol. 11 no. 3, p. 129-140.

Marceau D.J., Howarth P.J., Gratton D.J., Remote sensing and the measurement of geographical entities in a forested environment; part 1 The scale and spatial aggregation problem, *Remote Sensing of environment*, 1994, Vol. 49, N° 2, p. 93-104.

Mahfoud I., Josselin D., Fady B., *submitted,* Sensibilité des indices de (bio)diversité à l'effet d'agrégation, presented at SAGEO'2006, 11-13 septembre 2006, Strasbourg, 15 p.

Openshaw S., The modifiable areal unit problem. *Concepts and Techniques in Modern Geography*. 1984, Number 38, Geo Books, Norwich.

Rastetter E.B., King A.W., Cosby B.J., Hornberger G.M., O'Neill R.V., Hobbie J.E., « Aggregating fine-scale ecological knowledge to model coarser-scale attributes of ecosystems ». *Ecological Applications 2*, 1992, p. 55-70.

Reynolds, H. D., *« The modifiable areal unit problem: empirical analysis by statistical simulation »*. 1998, Thesis, University of Toronto.

Robinson A.H., « Ecological correlation and the behaviour of individuals ». *American Sociological Review,* 1950, N° 15, p. 1-357.

Wu J., Gao W., Tueller P.T., « Effects of changing spatial scale on the results of statistical analysis with landscape data: A case study », *Geographic Information Sciences 3*, 1997, p. 30-41.

Wu J., Levin S.A., « A spatial patch dynamic modelling approach to pattern and process in annual grassland ». *Ecological Monographs*, 1994, 64, p. 447-467.

Yule, G.U. and Kendall, M.G., *An introduction to the theory of statistics*, 1950, Griffin, London.