

Technical Note

Multivariate Analysis of Incomplete Mapped Data

Stéphane Dray, Nathalie Pettorelli, and Daniel Chessel
Laboratoire de Biométrie et Biologie Evolutive
Université Claude Bernard Lyon

Abstract

Classical multivariate analyses are based on matrix algebra and enable the analysis of a table containing measurements of a set of variables for a set of sites. Incomplete mapped data consist of measurements of a set of variables recorded for the same geographical region but for different zonal systems and with only a partial sampling of this zone. This kind of data cannot be analysed with usual multivariate methods because there is no common system of sites for all variables. We propose a new approach using GIS technology and NIPALS, an iterative multivariate method, to analyse the spatial patterns of this kind of data. Moreover, an extension of our method is that it can be used for areal interpolation purposes. We illustrate the method in analysing data concerning the distribution of roe deer weights over several years in a reserve.

1 Introduction

For several years, GIS has been recognized as the best tool to store and manage compilation of spatially referenced data. Analysis of these data is often made more difficult by the fact the areal units differ among various data sets. This incompatibility arises, for example, from the fact that data come from different sources or that studied areas change over time (Gregory 2002). This problem is well known by GIS users and can be solved with areal interpolation methods (Goodchild and Lam 1980), which transform data from one system of areal units (source zones) to another (target zones). Areal interpolation methods often “take the form of interpolating the data from the source regions to the intersections of the source and the target regions, and then combining these appropriately to infer the data for the target regions” (Bloom et al. 1996). This

Address for correspondence: Stéphane Dray, UMR CNRS 5558, Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard Lyon 1, 69622 Villeurbanne Cedex, France. E-mail: dray@biomserv.univ-lyon1.fr

implies that areal interpolation requires that each areal units system covers the whole study region. If this is not the case two problems arise: (1) when one or more source polygons do not intersect (at all) with the target polygons, data from these polygons will not be transferred to the target coverage, (2) when one or more target polygons is not intersected by a source polygon, no estimation can be made for this target polygon.

Multivariate analysis is a natural tool to summarise large data sets. Standard methods such as principal component analysis (Hotelling 1933) or spatially constrained methods such as local or global PCA (Thioulouse et al. 1995) are commonly used with GIS (Guisan et al. 1999, Kadmon and Danin 1997, Zhang and Selinus 1998) to identify and represent multivariate spatial structures. Multivariate analyses are based on matrix algebra (singular value decomposition) and data must be contained in a matrix where each column represents a variable and each row represents a site (e.g. Greenacre 1984). Performing multivariate analyses is then very restricting because it requires that all sites must be sampled for all variables. If it is not the case, there are two alternatives: (1) estimation of missing values, or (2) exclusion of variables and sites containing missing values. From these considerations, we can easily deduce that classical multivariate analysis cannot be used directly on compilation of data sets with different areal units. If each system of areal units covers the whole study zone, a primary step of areal interpolation can be performed and multivariate analysis is then applied on estimated values. For other cases (i.e. partial sampling of the study zone), it seems that there is no evident solution to perform multivariate analysis.

The purpose of this paper is the analysis of incomplete cartographic data. This deals with data collected for the same geographical region but for different zonal systems and with only a partial sampling of this zone. The analysis of spatial variations of this kind of data is not possible with usual multivariate analyses because the data cannot be entered in a variable-by-individual matrix. We propose a new methodology based on a joint use of GIS technology and multivariate analyses. We analyse a data set on the distribution of roe deer weights over years in the Chizé reserve (France) to illustrate the method. The sampling scheme adopted for this study involves that the sampling locations where the data have been collected are not considered as points (like for most studies on spatial data) but as polygons. Indeed, in a particular capture session, people fenced some given forestry plots with 2–5 km of nets and animals enclosed in this area were counted and weighed. Each year, from 58 to 79% of the area of the reserve is sampled and the sampling scheme (and so the sampling units) changes over years (Figure 1). In our example, variables correspond to different dates and so we analyse the temporal variations of a spatial structure.

2 Areal Interpolation and Changes in Support

2.1 Literature Review

It is common in geographical research that areal units for which data are available are not necessarily the ones the analyst wants to study. An answer to this problem can be obtained by areal interpolation (Goodchild and Lam 1980, Lam 1983) that consists of transferring data collected originally on one set of areal units (source zones) to a different set of areal units (target zones). This problem is well known by GIS users and various kinds of data transfer have been proposed in the literature. The “polygon overlay” (Markoff and Shapiro 1973) also referred to as “areal weighting” is probably the

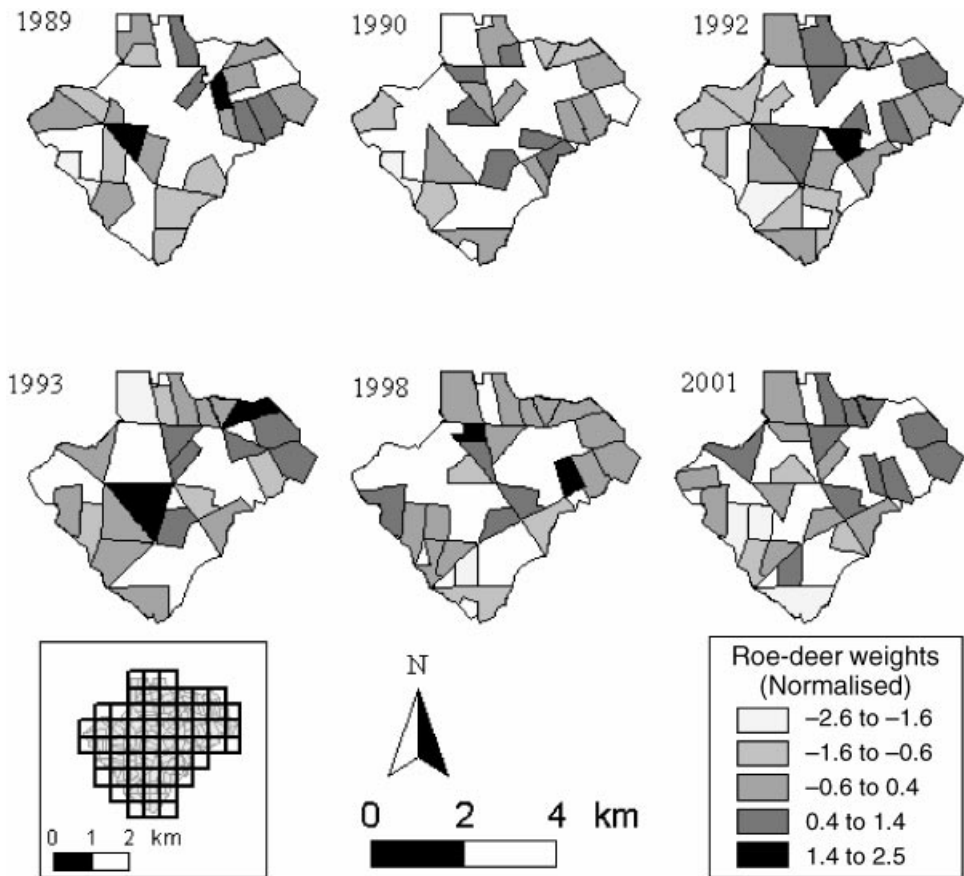


Figure 1 Sampling schemes and distribution of roe deer weights in the Chizé reserve. The number, size, shape and location of sampling areas are different among years. Weights have been normalised by year. The reference grid that has been chosen for the analysis is also represented on the map of forestry plots

simplest method and has become a major function of many GIS softwares (Bloom et al. 1996). Data for target zones are estimated as a weighted average (or weighted sum) of data for the source zones with which they intersect. Weighting is in proportion to the area of the zones of intersection. The “pynophylatic interpolation” originally suggested by Tobler (1979) improves the areal weighting method considering the values of neighbouring source zones. More “intelligent” methods, taking into account other relevant knowledge about the zones, have been developed. These methods are based on statistical assumptions providing maximum likelihood estimates of values for the target zones. They use the values of other ancillary variables to which the variable of interest may be related to improve the interpolation (Flowerdew and Green 1991, 1992, 1994, Flowerdew et al. 1991, Langford et al. 1991). Goodchild et al. (1993) present a general framework considering many of the previously mentioned methods.

These areal interpolation methods have been proposed in the case of two systems of areal units covering the entire study region. There is no doubt that these methods can

be easily extended to more than two zonal systems if one zonal system is chosen as the reference. Hence, multivariate analysis could be performed on real data for the reference zonal system and on estimation for other zonal systems. This approach is not satisfactory and introduces biases in the analysis because there are estimation errors for estimated data (Gregory 2002, Sadahiro 1999) and not for the real data of the reference areal units. Moreover, areal interpolation methods fail in the case of partial sampling because they cannot be used if there are target polygons with null intersections. To resolve these problems, we propose to operate changes in support by defining a reference layer of spatial units that covers the whole study zone, independently of the data. This allows preserving the symmetry of the data analysed and does not favour one system of areal units. We adopt an approach of areal weighting but other methods could be considered.

2.2 Spatial Linkage

We consider an area with defined boundaries. Some parts of this area are sampled at the first date (Figure 1). For the next dates (e.g. year 2, year 3 . . .), other sampled areas can be different or can overlap the previous sampled areas. The first step of our procedure is to create a reference layer of spatial units. Administrative units or some other kinds of space partitioning can be used to define the spatial units. In this paper, we chose to define the spatial units as the quadrats of a grid. The choice of the quadrat size is discussed below. Then, for each year, it is easy to construct neighbouring relationships between the quadrats of the reference grid and the sampling areas (Figure 2).

Let us consider, for year p , that k_p areas have been sampled. We construct a grid of n quadrats. The easiest way to establish a neighbouring relationship is to construct a matrix A_p with n rows and k_p columns where:

$$A_{p\,ij} = 1 \quad \text{if quadrat } i \text{ intersects the sampling area } j$$

$$A_{p\,ij} = 0 \quad \text{otherwise}$$

This neighbourhood matrix represents the strength of the potential interaction between quadrats and sampling areas. A more elegant and realistic way to fill this spatial weight matrix is to take into account the area of overlap between sampling areas and quadrats. The matrix A_p is then filled as follows:

$$A_{p\,ij} = \frac{S_{i \cap j}}{\sum_{j=1}^{k_p} S_{i \cap j}} \quad (1)$$

where $S_{i \cap j}$ is the area of the intersection between quadrat i and sampling area j . The data reproduced in Figure 2 can be used to illustrate this statement. The link between quadrat Q5 and sampling area P1 is simply expressed as $\frac{a_1}{a_2 + a_1}$ and as $\frac{a_2}{a_2 + a_1}$ for Q5 and P2.

2.3 Construction of the Data Table

Let us consider a quantitative variable X measured in all sampling areas for each date. Hence for each year p , data consist of a vector X_p with k_p rows (Figure 3). For each year, averages of X , weighted by overlapped area, can be computed for each quadrat

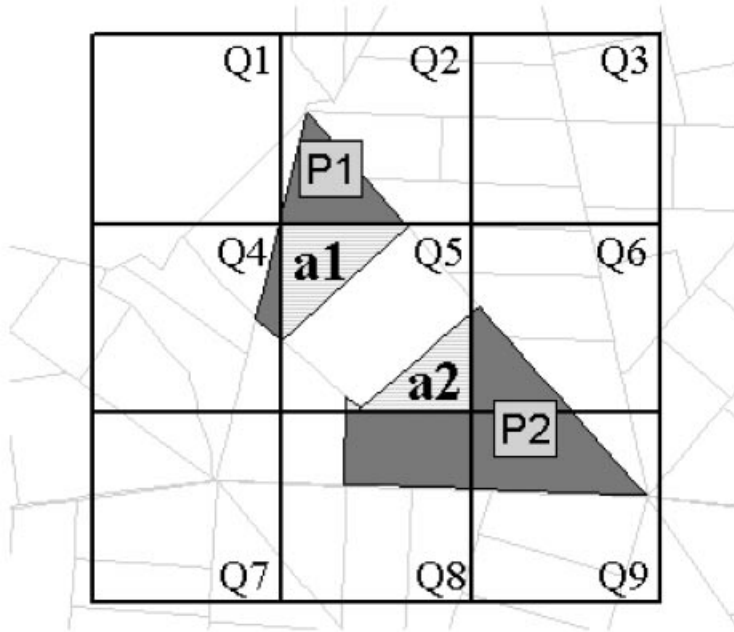


Figure 2 Crossing of the reference grid and sampling area. Establishment of the neighbourhood matrix between quadrats and sampling area is made by computing the area of the intersection between a quadrat and a sampling area. For example, the link between the quadrat Q5 and the sampling area P1 is simply expressed by $a1/(a1 + a2)$

and result in a vector Z_p with n rows. Computation of the weighted average for the i -th quadrat for year p (i.e. i -th element of Z_p) is simply the product of the i -th row of A_p by vector X_p . But, if for year p no sampling area intersects a quadrat i (i.e. sum of elements of the i -th row of A_p is null) then a missing value is assigned for the i -th element of vector Z_p . A matrix Z with n rows and N (total number of years) columns is then constructed in binding the N vectors Z_p . Applying classical multivariate analysis using singular value decomposition on table Z is not possible because of the existence of missing values. An alternative is the use of the NIPALS algorithm (Wold 1966) instead of singular value decomposition.

3 NIPALS Analysis

NIPALS (Nonlinear estimation by iterative partial least squares) is an algorithm, which is at the root of PLS regression. Wold (1966) presented this algorithm under the name of NILES (Nonlinear estimation by Iterative Least Square) in the case of PCA. NIPALS allows performing a PCA (principal component analysis) with missing values without deleting individuals with missing data or estimating the missing data. The algorithm is iterative and based on successive linear regressions (Tenenhaus 1998). A general presentation of the NIPALS algorithm is given in Wold et al. (1987). The method is shown in schematic form in Figure 3 and the algorithm used for NIPALS analysis with missing data is defined as follows:

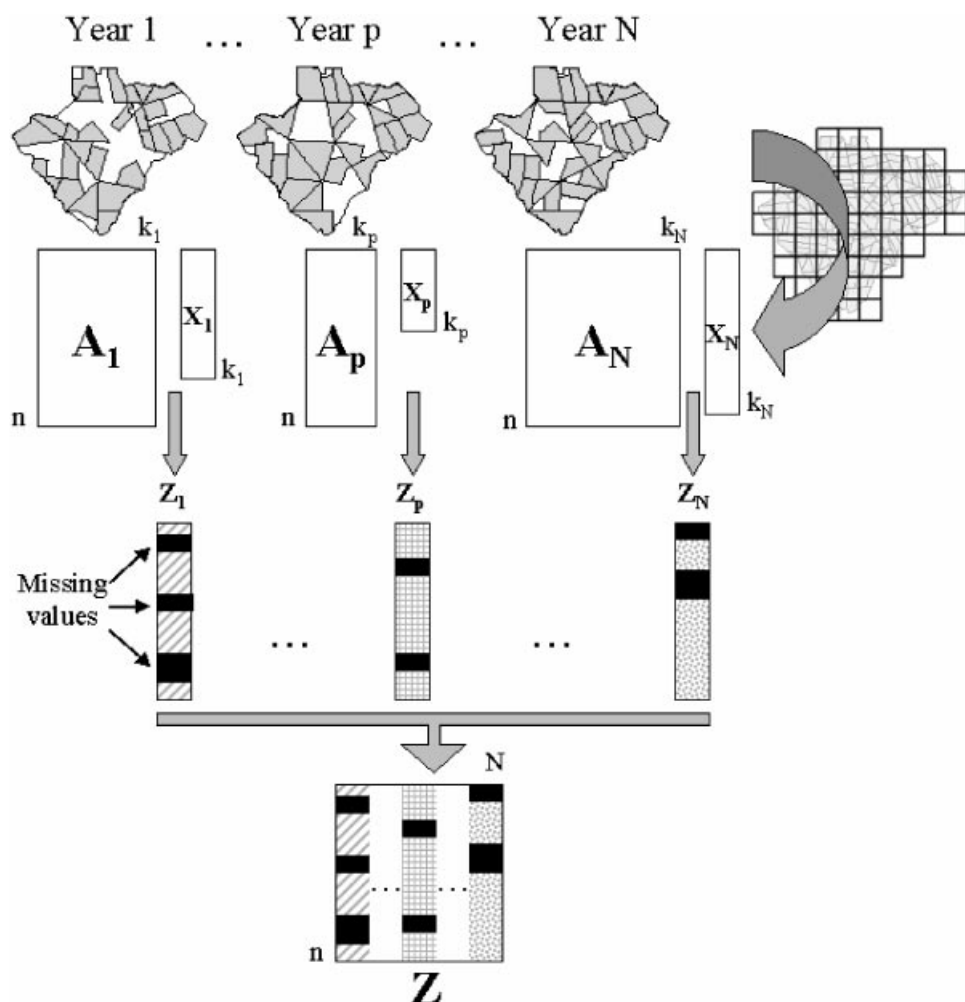


Figure 3 Schematic representation of the method. Crossing the sampling areas and the reference grid allows us to obtain a neighbourhood matrix (A_i) for each year i , measurements of variable (X_i) and neighbourhood matrix (A_i) are used to compute the average for each quadrat. A new table crossing quadrats and years is then created in binding all vectors Z_i . Table Z will be analysed using the NIPALS algorithm

Step 1: Normalisation of Z (Z_0)

Step 2: For $b = 1, 2, \dots, a$ (with $a \leq N$):

Step 2.1: $t_b = Z_{b-1}[1]$

Step 2.2: Repeat until the convergence of p_b

Step 2.2.1: For $j = 1, 2, \dots, N$:

$$p_b[j] = \frac{\sum_{i=1}^n \text{[if } Z[i, j] \text{ and } t_b[i] \text{ exist]} Z_{b-1}[i, j] t_b[i]}{\sum_{i=1}^n \text{[if } Z[i, j] \text{ and } t_b[i] \text{ exist]} t_b[i]^2}$$

Step 2.2.2: Normalise \mathbf{p}_b

Step 2.2.3: For $i = 1, 2, \dots, n$:

$$\mathbf{t}_b[i] = \frac{\sum_{j=1}^N \text{[if } \mathbf{Z}[i, j] \text{ exist]} \mathbf{Z}_{b-1}[i, j] \mathbf{p}_b[j]}{\sum_{j=1}^N \text{[if } \mathbf{Z}[i, j] \text{ exist]} \mathbf{p}_b[j]^2}$$

Step 2.3: $\mathbf{Z}_b = \mathbf{Z}_{b-1} - \mathbf{t}_b \mathbf{p}_b'$

In this way, NIPALS allows a user to perform PCA with missing values without estimating or deleting empty records. As for classical PCA, NIPALS allows the user to compute row (\mathbf{t}_b) and column (\mathbf{p}_b) coordinates as well as an eigenvalue for the b -th axis:

$$\lambda_b = \frac{1}{n-1} \mathbf{t}_b' \mathbf{t}_b \quad (2)$$

Moreover, missing values can be estimated using classical reconstitution formulae at the b -th order (Good 1969):

$$\hat{\mathbf{Z}}_0[i, j] = \sum_{l=1}^b \mathbf{p}_l[j] \mathbf{t}_l[i] \quad (3)$$

A PCA of table \mathbf{Z} with the NIPALS algorithm aims to find a row score that is a compromise of the different spatial patterns observed for all variables (i.e. years).

4 Application: Spatio-temporal Variation of Roe-deer Weights

This study was carried out in the 2,614 ha fenced Chizé reserve situated in western France (46°05'N, 0°25'W). The roe deer population has been intensively monitored by capture-mark-recapture methods since 1978 (Gaillard et al. 1993). Ten days of capture in January and February allow 150–350 roe deer to be caught each year. In a particular capture session, more than 100 people are involved and drive animals into 2–5 km of nets, enclosing some given forestry plots. Most animals are released with individual collars and the remainder is exported.

Fawns were captured between January and February and weighed using an electronic balance. The site of capture and the sex were noted. All information was transferred into a GIS. Sampling areas vary from one year to another (Figure 1). Male fawns are slightly heavier than female fawns (Gaillard et al. 1996), so adjusted weights for males were computed with an ANOVA in order to include in the analysis all individuals captured.

The reference layer was chosen as a grid of 58 quadrats measuring 800 m on a side. We chose this size to be consistent with the scale of the data because the area of a quadrat (0.64 km²) corresponds roughly to the average sampling area (0.654 km²). GIS was used to compute the table of weight means crossing the 58 quadrats and the 6 years. In this table, there is more than 6% missing values. The year 1998 was the poorest sampled (10% missing values) and three quadrats contained only values for four years out of the six total. Convergence was obtained in the NIPALS analysis and the decrease of eigenvalues suggests a one-axis structure (Figure 4a). All years are positively correlated with axis 1 (Figure 4b) which indicates that for these years heavy roe deer are found in the Northeast part of the reserve whereas light roe deer are found in the South (Figure 4c). However, coordinates of years on the correlation circle indicate that some

years (e.g. 1992) are more correlated to the common structure than others (e.g. 1998). Reconstitution formulae (Equation 3) based on the first axis of NIPALS analysis were used to obtain estimates of the weight distribution for all years for the 58 quadrats (Figure 4c). We used GIS to perform areal interpolation (i.e. areal weighting) to estimate weights for each zonal system (target zones) from the grid of quadrats (source zone). The results are satisfying if we consider that reconstitution of the data have been made using only one axis (Figure 5). The total sum of squares of differences between observed data and estimation is 222.57 (Table 1). There are obvious differences in fit between years and this can be confirmed by the values of correlation coefficients between observed data and estimation per year (Table 2). The best fit is obtained for 1992 and the worst for 1998. This is probably due to the fact that year 1992 is more correlated to the common structure identified on the first axis of NIPALS analysis than 1998 (Figure 4b). Moreover, 1998 is also correlated with the second axis and it is evident that the fit for 1998 will be greatly improved if the reconstitution is performed on the first two axes of NIPALS analysis.

Variations of quadrat size (Tables 1 and 2) influence NIPALS analysis. With the decrease of the size, the number of quadrats and the percentage of missing values increase. If the number of missing values increases, then the convergence is attained with difficulty and the number of iterations increases. Nevertheless, the use of small quadrats induces a finer-scale study and local patterns of variations can be detected. Therefore, the first eigenvalue increases and estimation is better adjusted with the detection of local patterns. This is confirmed by the fact that global correlation coefficients decrease as the quadrat size increases. Year 1992 is always the best fitted whereas lowest correlation coefficients were produced for years 1993, 1998, and 2001. This confirms results obtained for a size of 800 m and suggests a more sophisticated spatial structure than a one-axis structure for these years. In our example, the influence of the quadrat size on the results is minor because the structure observed in the roe-deer weights distribution is simple and strong and is detected easily at each spatial scale.

5 Discussion and Conclusions

Our approach requires that the user specify a reference layer of spatial units. These spatial units are the statistical individuals in the NIPALS analysis. The choice of this layer can be induced by the data but in most cases, the user must create this space partitioning and so the simplest way is to create a grid of quadrats. The choice of the size of quadrats has to be consistent with the spatial scale of the study. In our case, the area of quadrat is roughly the area of sampled plots. Moreover, the size of quadrats influences the number of missing values in the new data table. The smaller the quadrats, the larger the number of missing values is, because the number of intersections decreases. If there are too many missing values, convergence of NIPALS will not be attained. In the other way, the use of large quadrats will decrease the efficiency of the method to detect local structure and thereby the quality of the estimation. Nevertheless, it is obvious that the choice of the size has to be decided by considering the data a priori. In our case, sampling locations are polygons and so it is easy to establish the neighbourhood relationships by considering intersection across polygons. If the sampling locations are points, the use of buffer zones or more sophisticated methods such as tessellation (Green and Sibson 1978) can be used to assign a polygon to each sampling location and define neighbours.

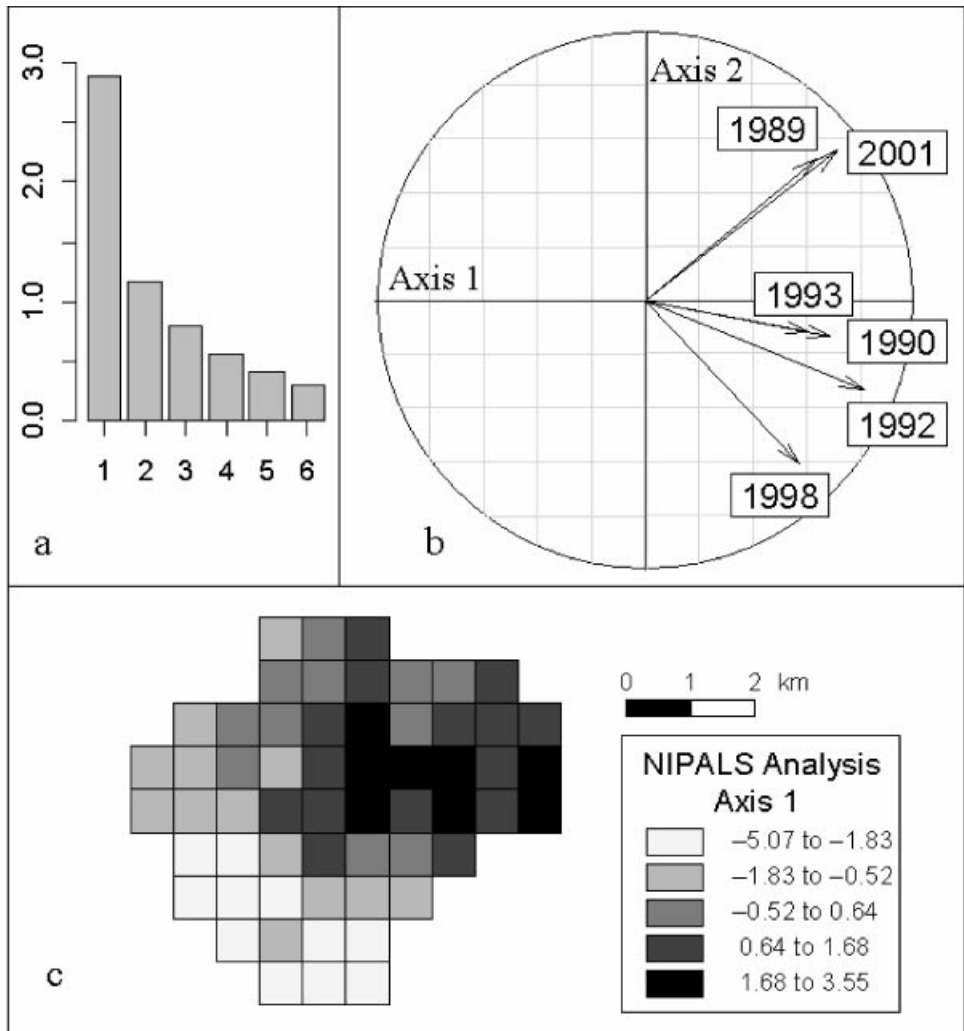


Figure 4 NIPALS analysis of roe deer weights data: (a) Eigenvalues, (b) Correlation circle, and (c) Spatial mapping of factor scores of the 58 quadrats for the first axis

Estimations of the data by reconstitution formulae are not very satisfactory but this is not surprising. Indeed, estimations have been made using only one axis and will be greatly improved with more axes. However, the first aim of our method is to perform multivariate analysis in the case of different zonal systems and partial sampling and not to predict data for new locations. This approach allows us to identify the most important structures in the data and to obtain information on spatial patterns for the whole study region from partial spatial data. Our method contains two steps of areal weighting and it is evident that it can be considered for areal interpolation purposes. Nevertheless, for this task, the estimations will be better with quadrats of small size and by using more axes in the reconstitution formula. Estimation of the data is made on the basis of the structures identified by the first axes of the analysis. Therefore, we estimate a year by taking into account the global structure of all years and considering only the common

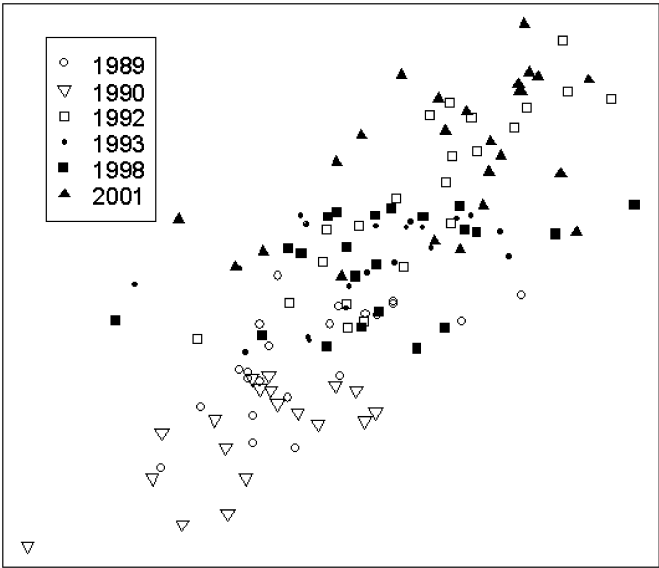


Figure 5 Estimated and observed roe deer weights. Estimates of weights are computed for each quadrat using reconstitution formulae after NIPALS analysis. Then, these estimates are used to compute the averages, weighted by area of intersection, for each sampled area. Correlations are given in Table 2

Table 1 NIPALS analysis of roe-deer weights data with quadrats of varying size. The total number of quadrats and the percentage of missing values is given for each quadrat size. Moreover, the first eigenvalue and the number of iterations from NIPALS analysis are noted. The error sum of squares (ESSQ) between observed data and estimation computed using first axis of NIPALS analysis and GIS are given

Quadrat size (m)	100	200	400	600	800	1000	1200
Number of quadrats	2819	739	203	99	58	41	30
% of missing values	33.1	25.4	16.2	10.6	6	4.4	4.4
Number of iterations	20	17	15	13	10	12	12
First eigenvalue	3.41	3.19	3.03	2.85	2.88	2.86	2.73
ESSQ	175.37	181.35	195.45	213.99	222.57	233.14	244.60

spatial variability among years. Hence, it is evident that the fit will be improved if two or more common structures were taken into account. In this context, the NIPALS approach can be considered as an intelligent areal interpolation method because for each variable, the estimation takes into account the values of other variables defining the NIPALS axes.

In this study, GIS appears as the central part of a multidisciplinary problem. GIS has been used to capture, manage and display the data. The representation of the data in the GIS has motivated new biological questions. The joint use of GIS and statistical analyses results in the elaboration of new methodology that helps with the resolution of

Table 2 NIPALS analysis of roe-deer weights data with quadrats of varying size. The correlation coefficients between observed data and estimates were obtained by areal interpolation of values obtained by reconstitution formulae on the first axis of NIPALS analysis. The number of areal units is given

Quadrat size (m)	100	200	400	600	800	1000	1200
1989 ($n = 22$)	0.87	0.85	0.80	0.76	0.67	0.70	0.59
1990 ($n = 19$)	0.79	0.76	0.77	0.67	0.69	0.68	0.67
1992 ($n = 22$)	0.92	0.91	0.89	0.85	0.87	0.80	0.77
1993 ($n = 21$)	0.53	0.52	0.49	0.49	0.51	0.50	0.39
1998 ($n = 22$)	0.59	0.55	0.46	0.42	0.42	0.36	0.37
2001 ($n = 25$)	0.57	0.60	0.62	0.59	0.54	0.53	0.53
All years ($n = 131$)	0.81	0.81	0.79	0.77	0.76	0.74	0.73

biological problems. So, the integration of GIS allows us to improve the statistical methodology as well as biological knowledge. GIS users get to know spatial analysis tools from geostatistics. Problems related to the integration of geostatistics in GIS softwares have been discussed for some time (e.g. Anselin and Getis 1992, Goodchild et al. 1992). These reflections have produced numerous packages linking geostatistics and GIS (Bivand 2001), and GIS and geostatistics are now considered rightly as essential partners for spatial analysis (Burrough 2001). In the same way, we think that the analysis of spatial data would probably benefit for the improvement of the links between GIS and multivariate analyses.

Acknowledgments

We thank the Office National de la Chasse for allowing us to use these data. We are grateful to all the students, field assistants, and volunteers that spent time catching and monitoring the roe deer fawns on the study site. Special thanks to Roger Bivand, Jean-Michel Gaillard and two anonymous reviewers for ideas, comments and suggestions on previous drafts of this paper.

References

- Anselin L and Getis A 1992 Spatial statistical analysis and geographic information systems. *Annals of Regional Science* 26: 19–33
- Bivand R 2001 More on spatial data analysis. *R News* 1: 13–7
- Bloom L M, Pedler P J, and Wragg G E 1996 Implementation of enhanced areal interpolation using MapInfo. *Computers and Geosciences* 22: 459–66
- Burrough P A 2001 GIS and geostatistics: Essential partners for spatial analysis. *Environmental and Ecological Statistics* 8: 361–77
- Flowerdew R and Green M 1991 Data integration: Statistical methods for transferring data between zonal systems. In Masser I and Blakemore M (eds) *Handling Geographical Information: Methodology and Potential Applications*. Harlow, Longman: 38–54
- Flowerdew R and Green M 1992 Developments in areal interpolation methods and GIS. *Annals of Regional Science* 26: 67–78

- Flowerdew R and Green M 1994 Areal interpolation and types of data. In Fotheringham A S and Rogerson P (eds) *Spatial Analysis and GIS*. London, Taylor and Francis: 121–45
- Flowerdew R, Green M, and Kehris E 1991 Using areal interpolation methods in geographic information systems. *Papers in Regional Science* 70: 303–15
- Gaillard J M, Delorme D, Boutin J M, Van Laere G, and Boisaubert B 1996 Body mass of roe deer fawns during winter in two contrasting populations. *Journal of Wildlife Management* 60: 29–36
- Gaillard J M, Delorme D, Boutin J M, Van Laere G, Boisaubert B, and Pradel R 1993 Roe deer survival patterns: A comparative analysis of contrasting populations. *Journal of Animal Ecology* 62: 778–91
- Good I J 1969 Some applications of the singular decomposition of a matrix. *Technometrics* 11: 823–31
- Goodchild M, Anselin L, and Deichmann U 1993 A framework for the areal interpolation of socioeconomic data. *Environment and Planning A* 25: 383–97
- Goodchild M, Haining R, and Wise S 1992 Integrating GIS and spatial data analysis: Problems and possibilities. *International Journal of Geographical Information Systems* 6: 407–23
- Goodchild M and Lam N S 1980 Areal interpolation: A variant of the traditional spatial problem. *Geoprocessing* 1: 297–312
- Green P and Sibson R 1978 Computing Dirichlet tessellations in the plane. *The Computer Journal* 21: 168–73
- Greenacre M J 1984 *Theory and Applications of Correspondence Analysis*. London, Academic Press
- Gregory I N 2002 The accuracy of areal interpolation techniques: Standardising 19th and 20th century census data to allow long-term comparisons. *Computers, Environment and Urban Systems* 26: 293–314
- Guisan A, Weiss S B, and Weiss A D 1999 GLM versus CCA spatial modeling of plant species distribution. *Plant Ecology* 143: 107–22
- Hotelling H 1933 Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* 24: 417–41
- Kadmon R and Danin A 1997 Floristic variation in Israel: A GIS analysis. *Flora* 192: 341–45
- Lam N S 1983 Spatial interpolation methods: A review. *The American Cartographer* 10: 129–49
- Langford M, Maguire D J, and Unwin D J 1991 The areal interpolation problem: Estimating population using remote sensing in a GIS framework. In Masser I and Blakemore M (eds) *Handling Geographical Information: Methodology and Potential Applications*. Harlow, Longman: 55–77
- Markoff J and Shapiro G 1973 The linkage of data describing overlapping geographical units. *Historical Methods Newsletter* 7: 34–46
- Sadahiro Y 1999 Accuracy of areal interpolation: A comparison of alternative methods. *Journal of Geographical Systems* 1: 323–46
- Tenenhaus M 1998 *La régression PLS. Théorie et Pratique*. Paris, Editions Technip
- Thioulouse J, Chessel D, and Champely S 1995 Multivariate analysis of spatial patterns: A unified approach to local and global structures. *Environmental and Ecological Statistics* 2: 1–14
- Tobler W R 1979 Smooth pycnophylactic interpolation for geographical regions. *Journal of the American Statistical Association* 74: 519–36
- Wold H 1966 Estimation of principal components and related models by iterative least squares. In Krishnaiah P R (eds) *Multivariate Analysis*. New York, Academic Press: 391–420
- Wold H, Esbensen K, and Geladi P 1987 Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 2: 37–52
- Zhang C S and Selinus O 1998 Statistics and GIS in environmental geochemistry: Some problems and solutions. *Journal of Geochemical Exploration* 64: 339–54