DATA ANALYSIS AND DATA MINING GROUP

Tomas Aluja
Departament d'Estadística i Investigació Operativa
Universitat Politècnica de Catalunya

Michael Greenacre
Departament d'Estadística
Universitat Pompeu Fabra

For some time there has been a strong cooperation with the French group of «Analyse des Données», inspired by the seminal work of J.-P. Benzécri. This cooperation has led to continuous research exchanges centred in the Universitat Politècnia de Catalunya, Universitat Pompeu Fabra and Universitat de Girona, among others.

Data Analysis in this context refers to the conception of statistics based on data, without any probabilistic assumptions, building models that fit the data, instead of the classical approach of using the data to validate hypothesized models within a probabilistic data framework. When applied to large data sets, this approach is often referred to as Data Mining (Aluja, 2001).

The research has focused on methodological innovation as well as software implementations, some of them commercial, of the proposed methodologies.

Three international conferences have organized on this theme, jointly with the Central Archive for Empirical Social Research of the University of Cologne in Germany: in 1991, Correspondence Analysis in the Social Sciences; in 1995, Visualization of Categorical Data; and in 1999, Large Scale Data Analysis. Two of these conferences led to edited books (Greenacre and Blasius, 1994; Blasius and Greenacre, 1998). The next conference in this series is Correspondence Analysis and Related Methods, taking place in 2003 in Barcelona.

The following are topics of research by this group.

Detection of functional regions: This is a classical problem in regional statistics, a functional region being defined as a pole of attraction with its hinterland. It was applied to the case of the «comarcal» (county) delimitation of Catalonia. A new logical distance was used to measure the attraction between municipalities (Aluja, 1983).

Local principal components analysis: Very often, principal component analysis reveals a previously known global variability. In such cases it is interesting to go beyond the classical analysis, eliminating the known variability in the data. This has been proposed under different names and approaches. Following previous work by Ludovic Lebart, Aluja *et al.* (1985) and Aluja (1988) developed a non-parametric approach using a non-oriented graph on the individuals. This is a very flexible tool allowing to partial out

the variability «modelled» by the edges of the graph. It is possible to obtain a triplot of the local variables, the edges of the graph and the individuals after having eliminated the unwanted effect. This methodology was extended to the analysis of panel data by Aluja *et al.* (1993). Some equivalences between the different approaches are presented by Nonell, Thió and Aluja (2000).

Decision trees: The objective of decision trees methods is to automatically detect which variables serve to explain the behaviour of a given response variable, either quantitative or categorical. Our research focused on the problem of the stability of tree. The tree-growing process is highly dependent on the data, especially in actual data mining applications which have a large number of variables, where the construction of decision trees is somewhat arbitrary since we have to select between several splits possessing similar explanatory powers. Tackling this problem, Aluja and Nafria (1998*a*) proposed a new general geometric formulation of the impurity of a node, which allows us to compute the contribution of each individual to the impurity, and hence to detect influential individuals, and also defined an alternative criterion for the classification of trees based on a generalized Smirnov distance. In addition, Aluja and Nafria (1998*c*) studied the complexity of the CART methodology, proposing an algorithm with linear cost with the number of individuals. A complementary research line was to incorporate expert knowledge into the clustering process (Gibert, Aluja and Cortes, 1998).

Data fusion and grafting: Data fusion refers to the problem of merging information coming from independent sources. It is also known as statistical matching. The fusion consists of the transferring the specific y variables from the donor file to the receptor. There are two main methodologies for data fusion, one is modelling the relation of the y variables respect to the common x variables in the donor file and apply it to the receptor file. Thye other approach is the so-called «hot deck» methodology, which involves finding for each individual of the recipient file one or more similar individuals in the donor file and then transferring the y values of these individuals to the receptor. Our research is to assess the viability of the fusion process, to compare the accuracy of the different methodologies proposed, and to propose validation measures of the process (Aluja *et al.*, 1998*b*, Aluja and Thió, 2001) as well as implementing this in a complete system for data fusion. File grafting is a related methodology complementing the previous research line, developed to visually display information coming from independent sources in a common factorial subspace (Aluja, Morineau and Rius, 1999; Rius, Aluja and Nonell, 1999). Grafting is implemented within the SPAD software.

Textual data analysis: The usual techniques of multivariate data analysis can be applied to textual data, which form a homogeneous corpus full of redundancy. Correspondence analysis can be used to associate the lexical information with any external information about texts. The research has focused on the problem of definition of complex lexical units (repeated segments) to take into account the context in the analysis, the application of such methodology to massive corpus of textual data and the simul-

taneous analysis of different open questions (Bécue, 1993, 1997*a*, Bécue and Lebart, 1998, 2000) and the analysis of questions in different languages (Bécue, 1997*b*). A new methodology, Multiple Factorial Analysis, has been proposed to analyze three-way contingency tables, either textual or numerical (Bécue, 1998, Bécue and Pagès, 2001). The proposed methodologies have been implemented in the software system SPAD-textual.

Multivariate methods used in environmental research: This research line concerns the study of multivariate methods that are used in environmental studies. Important multivariate methods in the environmental sciences are principal component analysis (PCA), factor analysis, correspondence analysis (CA) and methods using linear constraints such as canonical correspondence analysis (CCA) and redundancy analysis (RDA). CCA and another approach based on multidimensional scaling are compared by Greenacre and Fieler (1995). Interesting applications of these methods have been obtained with a large database of marine biological samples. Quality statistics for the goodness of fit of all data matrices involved in CCA have been derived (Graffelman 2001*a*). Research has focused on the representation of supplementary information (cases and variables) in biplots obtained by these methods, with attention to the different types of scaling and the geometrical properties of the solution. Some theoretical results are given by Graffelman (1999, 2000, 2001) and Graffelman and van der Velden (1999). A different line of research undertaken involves various biometric analyses of the human sex ratio (Graffelman *et al.*, 1999), Graffelman and Hoekstra, 2000).

Correspondence analysis: Research here has focused on ways of interpreting multiple correspondence analysis and on a different algorithm for fitting multiway contingency tables called «joint correspondence analysis» (Greenacre, 1993*b*, 1994*a*, 1994*b*, 1998). In joint correspondence analysis the diagonal subtables on the super-diagonal of the so-called Burt matrix are not fitted by least-squares, only the extra-diagonal tables. This leads to correct measures of explained inertia in correspondence analysis. As a by-product, the usual solution in multiple correspondence analysis can be improved by a simple calculation which provides a lower bound, and usually a close lower bound, to the correct percentages of inertia (see, for example, Greenacre, 1993*a*).

Biplots and unfolding: Biplots arising from correspondence analysis and multiple correspondence analysis are also studied (Greenacre, 1993*c*) and a comparison with biplots for compositional data is given by Aitchison and Greenacre (2002). Gower and Greenacre (1996) showed how unfolding can be applied to a square symmetric matrix of distances, leading to a novel way of interpreting a multidimensional scaling display. Visualization of preference data in a marketing context is given by Torres and Greenacre (2002).

Analysis of square asymmetric tables and matched matrices: In this line of research the joint visualization of two or more tables is studied. The analysis of a square asymmetric table is studied, the two tables being the original table and its transpose (Gree-

nacre, 2000). In the case of two rectangular tables, matched by rows and columns, their joint visualization can be achieved by the singular value decomposition of a block matrix where each matrix appears twice (Greenacre, 2001). The case of two square asymmetric matrices, usually transition matrices, is given by Greenacre and Clavel (2002).

References

1. Aitchison, J. and Greenacre, M. J. (2002). «Biplots of Compositional Data». *Applied Statistics*, 51, 375-392.
2. Aluja, T. (1983). «Determinació dels centres d'atracció i llur zona d'influència en funció dels equipaments municipals».In *El Debat de la Divisió Territorial de Catalunya. Edició d'estudis propostes i documents (1939-1983)*, Oriol Nel·lo and Enric Lluch (eds.), Altafulla, Barcelona, 839-863.
3. Aluja, T. and Lebart, L. (1985). «Factorial analysis upon a graph». *Demandez le Programme*, 3, 3-34, CISIA, París.
4. Aluja, T. (1988). «Local and partial correspondence analysis. Application to the analysis of electoral data». *Computational Statistics Quarterly*, 4, 89-103.
5. Aluja, T. and Nonell, R. (1991). «Local principal components analysis». *Qüestiió*, 15, 267-278.
6. Aluja, T. and Nonell, R. (1993). «Multiple correspondence analysis on panel data». In *Multivariate Analysis: Future Directions 2*, C. M. Cuadras and C. R. Rao (eds.), North-Holland, Amsterdam, 233-244.
7. Aluja, T. (1994). «Formation of an index of economic capacity in Barcelona». *Qüestiió*, 18, 377-384.
8. Aluja, T. and Nafria, E. (1998*a*). «Generalized impurity measures and data diagnostics in decision trees». In *Visualisation of Categorical Data*, Jörg Blasius and Michael Greenacre (eds.), Academic Press, San Diego, 59-70.
9. Aluja, T., Nonell, R., Rius, R. and Martinez-Abarca, M. J. (1998*b*). «Data fusion and file grafting». *Analyses Multidimensionelles des Données*, 7-14.
10. Aluja, T. and Nafria, E. (1998*c*). «Robust impurity measures in decision trees». In *Data Science, Classification and Related Methods*, Springer Verlag, Heidelberg, 207-214.
11. Aluja, T., Morineau, A. and Rius, R. (1999). «La greffe de fichiers et ses conditions d'application. Méthode et exemple». In *Enquêtes et Sondages. Méthodes, Modèles, Applications, Nouvelles Approches*. Dunod, París, 94-102.
12. Aluja, T. and Thió, S. (2001). «Survey data fusion». *Bulletin de Méthodologie Sociologique*, 72, 20-36.
13. Aluja, T. (2001). «La minería de datos, entre la estadística y la inteligencia artificial». *Qüestiió*, 25, 479-498.
14. Álvarez, R., Bécue, M. and Lanero, J. J. (2000). «Le vocabulaire gouvernemental espagnol (1979-1996)». *Mots*, 62, 31-47.

15. Bécue, M. (1993). «Utilisation des logiciels de statistique: la question ouverte de l'enquête». *Modulad*, 11, 50-58.
16. Bécue, M. (1997*a*). «Visualization of open questions: a French study of pupils' attitudes to Mathematics».In *Visualization of Categorical Data*, Jörg Blasius and Michael Greenacre (eds.), Academic Press, San Diego, 151-158.
17. Bécue, M. (1997*b*). «Etude comparative de réponses ouvertes à différentes questions». In *Analyses Multidimensionnelles des Données*, Fernández Aguirre, K. and Morineau, A. (eds.), Cisia, París, 65-72.
18. Bécue, M. (1998). «Three-way textual data analysis». In *Advances in Data Science and Classification, Springer Series of Studies in Classification, Data Analysis and Knowledge Organization*, Rizzi, A., Vichi, M. and Bock, H. H. (eds.), Springer Verlag, Heidelberg, 457-464.
19. Bécue, M. and Lebart L. (1998). «Clustering of texts using semantic graphs. Application to open-ended questions». In *Surveys. Data Science, Classification and Related Methods, Springer Series of Studies in Classification, Data Analysis and Knowledge Organization*, Hayashi, C., Oshumi, N., Yajima, K., Tanaba, Y., Bock, H. H. and Baba, Y. (eds.), Springer Verlag, Tokyo, 480-487.
20. Bécue, M. and Lebart, L. (2000). «Analyse statistique des réponses ouvertes. Application à des enquêtes auprès de lycéens». In *Analyse des Correspondances et Techniques Connexes*, J. Moreau, P. A. Doudin and P. Cazes (eds.), 59-83.
21. Bécue. M. and Pagès, J. (2001). «Analyse simultanée de questions ouvertes et de questions fermées. Méthodologie, exemple». *Journal de la Société Statistique de France*, 142-4.
22. Blasius, J. and Greenacre, M. J. (1994). «Computation of correspondence analysis». In Greenacre, M. J. and Blasius, J. (eds.), *Correspondence Analysis in the Social Sciences*, Academic Press, London, 53-78.
23. Blasius, J. and Greenacre, M. J. (1998). *Visualization of Categorical Data*, Academic Press, New York.
24. Gibert, K., Aluja, T. and Cortés, U. (1998). «Knowledge discovery with clustering based on rules. Interpreting results». In *Principles of Data Mining and Knowledge Discovery*, Springer Verlag, Berlin, 83-93.
25. Gower, J. C. and Greenacre, M. J. (1996). «Unfolding a symmetric matrix». *Journal of Classification*, 13, 81-105.
26. Graffelman, J. (1998). «Solution 97.5.3: A fundamental matrix result on scaling in multivariate analysis. Solution 97.5.3». *Econometric Theory*, 14, 693-694.
27. Graffelman, J. (1999). «Solution 99.1.5: The justification of multidimensional scaling under Euclidean conditions. Solution». *Econometric Theory*, 15, 908-909.
28. Graffelman, J. (2000). «Solution 99.5.1: Use of the Moore-Penrose Inverse in Canonical Correspondence Analysis». *Econometric Theory*, 16, 792-793.
29. Graffelman, J. and van de Velden, M. (1999). «Problem 99.4.4: Upper bounds for the eigenvalues of the product of a symmetric idempotent and a non-negative definite matrix». *Econometric Theory*, 15, 631.

30. Graffelman, J., Fugger, E. F., Keyvanfar, K. and Schulman, J. D. (1999). «Human live birth and sperm sex ratio compared». *Human Reproduction*, 14, 2917-2920.
31. Graffelman, J. and Hoekstra, R. F. (2000). «A statistical analysis of the effect of warfare on the human secondary sex ratio». *Human Biology*, 72, 433-445.
32. Graffelman, J. (2001*a*). «Quality statistics in canonical correspondence analysis». *Environmetrics*, 12, 485-497.
33. Graffelman, J. (2001*b*). «Factor Analysis». In El-Shaarawi, A. H. and Piegorsch, W. W. (eds), *Encyclopedia of Environmetrics. Volume 2*, John Wiley, Chichester, 763-767.
34. Greenacre, M. J. (1993*a*). *Correspondence Analysis in Practice*. Academic Press, London.
35. Greenacre, M. J. (1993*b*). «Different geometric approaches to correspondence analysis of multivariate data». In Opitz, O., Lausen, B. and Klar, R. (eds.), *Information and Classification*, Springer-Verlag, Berlin, 190-200.
36. Greenacre, M. J. (1993*c*). «Biplots in correspondence analysis». *Journal of Applied Statistics*, 20, 251-269.
37. Greenacre, M. J. (1994*a*). «Multiple and joint correspondence analysis». In Greenacre, M. J. and Blasius, J. (eds.), *Correspondence Analysis in the Social Sciences*, Academic Press, London, 141-161.
38. Greenacre, M. J. (1994*b*). «Correspondence analysis and its interpretation». In Greenacre, M. J. and Blasius, J. (eds.), *Correspondence Analysis in the Social Sciences*, Academic Press, London, 3-22.
39. Greenacre, M. J. (1995). «Multivariate generalisations of correspondence analysis». In Cuadras, C. M. and Rao, C. R. (eds.), *Multivariate Analysis: Future Directions 2*, North Holland, Amsterdam, 327-340.
40. Greenacre, M. J. (1998). «Diagnostics for joint displays in correspondence analysis». In Blasius, J. and Greenacre, M. J. (eds.), *Visualization of Categorical Data*, Academic Press, New York, 221-238.
41. Greenacre, M. J. (2000). «Correspondence Analysis of Square Asymmetric Matrices». *Applied Statistics*, 49, 297-310.
42. Greenacre, M. J. (2001). «Analysis of matched matrices». Working Report 539, Department of Economics and Business, Universitat Pompeu Fabra, submitted for publication.
43. Greenacre, M. J. and Blasius, J. (1994). *Correspondence Analysis in the Social Sciences*, Academic Press, London, 53-78.
44. Greenacre, M. J. and Clavel, J. G. (2002). «Simultaneous visualization of two transition tables». *Polish Journal of Statistics, Special Issue: Statistics in Transition*, in press.
45. Lebart, L., Morineau, A., Pleuvret, P. and Aluja, T. (1983). «Manuel SPAD. Système Portable pour l'Analyse des Données». CISIA, París.
46. Lebart, L., Morineau, A., Bécue, M. and Häusler, L. (1993). «Système pour l'Analyse de Données Textuelles, Manuel de l'Utilisateur». CISIA, París.

47. Nonell, R., Thió, S. and Aluja, T. (2000). «Some alternatives for conditional principal component analysis». *Applied Stochastic Models in Business and Industry*, 16, 189-200.
48. Rius, R., Aluja, T. and Nonell, R. (1999). «File grafting in market research». *Applied Stochastic Models in Business and Industry*, 15, 451-460.
49. Sieber, T. N., Petrini, O. and Greenacre, M. J. (1998). «Correspondence analysis as a tool in fungal taxonomy». *Systematic and Applied Microbiology*, 21, 433-41.
50. Torres, A. and Greenacre, M. J. (2002). «Dual scaling and correspondence analysis of preferences, ratings and paired comparisons». *International Journal for Research in Marketing*, to appear.