Statistics and Computing **3** (1993) 197–199

Real data are messy

PIERRE LEGENDRE

Département de sciences biologiques, Université de Montréal, C.P. 6128, Succ. A, Montréal, Québec, Canada H3C 3J7

Trying to predict what the future holds is certainly the best way to prove oneself wrong. Fifteen years ago, who would have predicted the present domination of microcomputers, the development of computer-intensive statistical methods of data analysis, and the world-wide e-mail communication networks? These were not trends then; the ideas behind these realizations have been developed by small groups of inventive people. The future certainly holds new methods based on revolutionary ideas that we cannot foresee at the present moment, as well as computing capabilities and working habits unknown at this time. So I will limit this note to a description of some important problems that, in my opinion, are worthy of statisticians' attention. I will speak of data analysis, the field comprising exploratory as well as confirmatory methods directed towards real problems and data.

Traditionally in fields of application such as biology, the teaching of basic and advanced statistics revolved around a series of questions and situations that were legitimate because answers were available in the realm of traditional parametric statistics. Data were assumed to be well behaved, observations were independent from one another (i.e. not autocorrelated), distributions were normal, samples were representative, missing values were non-existent or, in any case, the fault of negligent field people. Scientists studying natural phenomena, and among them ecologists, feel the need for a statistics capable of analysing the real, messy data they obtain every day. In ecology for instance, variables are increasingly semi-quantitative or qualitative in nature; when studying a diversity of environments, species abundances are plagued with large percentages of zeros; variables are autocorrelated spatially and temporally, while species analysed in community-level studies are phylogenetically autocorrelated; missing values do occur for a variety of legitimate reasons; sampling is not necessarily random. Scientists often turn to statisticians working in application fields such as geo-statistics or geography, where these problems are known, understood and respected, or even to non-statisticians who have used their imagination to develop empirical methods of analysis adapted to their needs. It is time for trained statisticians to get their hands dirty with real data. Here are a few interesting avenues.

Statistical methods

Is it possible to develop univariate and multivariate inferential statistical methods that are valid in the presence of dependencies of various kinds among the observations? I refer to spatial (Cliff and Ord, 1981), temporal, and phylogenetic (Harvey and Pagel, 1991) autocorrelation. Quite a bit of work has been done since the 1950s on Anova and the related discriminant analysis problem, especially in time series (as summarized by Crowder and Hand, 1990), and to a lesser extent on spatially autocorrelated data (mostly for regular grids, assuming an AR(1) generating process, or in the framework of experimental designs). Some work has also been done on goodness-of-fit tests, and on the tests of significance used in regression and correlation analysis (for example, Clifford *et al.* (1989) and Dutilleul (1993) have shown how to modify the t-test for assessing the correlation between two spatial processes.) But altogether, new parametric tests of significance have to be developed for most situations; modifications may imply calculating a modified number of degrees of freedom, an effective sample size or a modified estimate of the statistics' variance; another avenue is to design new permutational tests that leave the autocorrelation structure unmodified (an example is Legendre *et al.* 1990).

Sampling

In fields of application such as ecology, economics, epidemiology, genetics, geography, geology, marketing, political science, and sociology, sampling is traditionally assumed to follow the survey rules worked out by statisticians about 50 years ago. Representative samples are obtained by sampling at random from a statistical population, and geographical locations are assumed to be independent of one another. Textbooks often claim that random sampling ensures the validity of classical statistical tests, as if autocorrelation were to disappear from the data. Building on what is now known about spatial autocorrelation and its influence on statistical tests, and on spatial statistics and geostatistics, what guidelines could be given to practitioners about sampling design and subsequent analysis of the data?

Ordination methods

Classical ordination methods such as principal components analysis, correspondence analysis, and metric (principal coordinates analysis) or nonmetric scaling, allow to investigate the structure of multidimensional data. Constrained ordination methods such as redundancy analysis (van der Wollenberg, 1977) and canonical correspondence analysis (ter Braak, 1986) are generalizations of multiple regression analysis to multivariate dependent data tables. The 'partial' form of these methods allows the joint analysis of three multivariate data tables. It is possible to further extend this family of methods to analyse k data tables simultaneously? Could that lead to causal analysis ('path analysis') to test hypotheses involving those k tables as distinct logical entities?

Dissimilarity matrices

A large body of literature deals with dissimilarity matrices computed from univariate or multivariate data; an international conference, *Distancia* '92, was held last year in Rennes, France, on this subject. Such matrices represent a convenient way to transform ordinary data and make them comparable to data that naturally occur as dissimilarities, such as those found in the biochemical study of evolution (DNA or RNA hybridization or 'restriction fragment' data), behavioural studies (dyadic data, network data), or geographic distances among localities on maps. Matrices of this type can be compared using either the Mantel test of matrix correspondence (Mantel, 1967) or derived methods such as partial Mantel tests; a comparison of such methods is presented by Oden and Sokal (1992).

In many cases, dissimilarities in two or several matrices may not be in linear relation to one another, so that the linear statistics such as those used in the above-mentioned methods may be inefficient. How should one deal with these non-linearities? Present proposals involve data transformations (splines, ACE, LOWESS) or the use of non-linear statistics.

Proposals have been made (Hubert and Golledge, 1981; Smouse *et al.* 1986; Manly, 1986; Krackhardt, 1988) to generalize this approach and use k independent resemblance matrices as predictors in a multiple-regression-type framework. Again, could these regression coefficients be used in a causal analysis ('path analysis' on distance matrices: Leduc *et al.*, 1992) to test hypotheses involving those matrices as distinct logical entities?

Comparative studies

Between a canonical correlation approach (based on raw data tables) and a Mantel-type matrix approach, which one is preferable when the data could be subjected to both? Do they lead to the same or different results? Comparative studies are needed from both the theoretical and empirical points of view, using real as well as simulated data.

Structure in data sets

Before carrying out a cluster analysis, can we assess whether there is structure in a data set. Tests can be conceived based upon the raw data matrix or upon a derived resemblance matrix. Very few methods are available to do this, as can be seen in the recent review by Gordon (1993).

We also need good tests to determine the number of clusters that can be extracted from a data set; a plethora of indices have been proposed for this purpose, and compared by Milligan and Cooper (1985); yet, the sampling theory and distributions of these indices have not been properly worked out.

Is there a way to perform a goodness-of-fit test of the data to a specified grouping, when the grouping has been derived from the same data set? Perruchet (1983) wrote an early review of this subject; another approach is that of Legendre *et al.* (1984).

Computing

A good many of the classification and ordination problems are NP-hard; the k-means clustering and non-metric multidimensional scaling problems are examples, as are the problems of constructing evolutionary trees either from 'ordinary' variables or from molecular data. Could computer scientists develop polynomial-time algorithms? Or, at least, non-polynomial ones that remain efficient in practice to compute solutions to these important tasks? Could they provide proofs of the optimality of these algorithms?

Statistical software

Many of the statistical packages used by scientists in the application fields do not offer advanced solutions to the problems posed by messy data (autocorrelated data, missing values, etc.) When is the transfer of technology going to take place from the basic statistical literature to the person(s) or group(s) who develop these packages? This problem is less serious in advanced software used by statisticians, such as S+, SC, SPIDA, etc. Another example is inference from samples that have been taken from a geographic surface in a nonrandom fashion; this problem is discussed at some length in the geostatistical literature, but again these methods are not integrated into major packages.

How good is the commercially available statistical software used by the large majority of scientists in application fields? For instance, about the methods of cluster analysis offered by packages like SPSS, SAS, BMDP, or Systat: are they state-of-the-art, or one generation too old? What should we expect software companies to offer researchers in application fields in 1993? Software reviews that appear in journals like *The American Statistician, Applied Statistics*, or *Statistics and Computing* are a good way for knowledgeable statisticians to influence practices in the fields of application.

References

Cliff, A. D. and Ord, J. K. (1981) *Spatial processes: Models and applications*. Pion, London.

Clifford, P., Richardson, S. and Hémon, D. (1989) Assessing the significance of the correlation between two spatial processes. *Biometrics* **45**, 123–134.

Crowder, M. J. and Hand, D. J. (1990) *Analysis of Repeated Measures*. Chapman and Hall, London.

Dutilleul, P. (1993) Modifying the *t* test for assessing the correlation between two spatial processes. *Biometrics*, (in press).

Gordon, A. D. (1993) Hierarchical classification. In *Clustering and Classification*. (P. Arabie, L. Hubert and G. De Soete, eds.) World Scientific, River Edge, NJ. To appear.

Harvey, P. H. and Pagel, M. D. (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford.

Hubert, L. J. and Golledge, R. G. (1981) A heuristic method for the comparison of related structures. *Journal of Mathematical Psychology*, **23**, 214–226.

Krackhardt, D. (1988) Predicting with networks: nonparametric multiple regression analysis of dyadic data. *Social Networks*, **10**, 359–381.

Leduc, A., Drapeau, P., Bergeron, Y. and Legendre, P. (1992) Study of spatial components of forest cover using partial Mantel tests and path analysis. *Journal of Vegetation Science*, **3**, 69–78.

Legendre, P., Dallot, S. and Legendre, L. (1984) Succession of species within a community: chronological clustering, with applications to marine and freshwater zooplankton. *The American Naturalist*, **125**, 257–288.

Legendre, P., Oden, N. L., Sokal, R. R., Vaudor, A. and Kim, J. (1990) Approximate analysis of variance of spatially autocorrelated regional data. *Journal of Classification*, **7**, 53–75.

Manly, B. F. J. (1986) Randomization and regression methods for testing for associations with geographical, environmental, and biological distances between populations. *Research in Population Ecology*, **28**, 201–218.

Mantel, N. (1967) The detection of disease clustering and a generalized regression approach. *Cancer Research*, **27**, 209–220.

Milligan, G. W. and Cooper, M. C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.

Oden N. L. and Sokal, R. R. (1992) An investigation of three-matrix permutation tests. *Journal of Classification*, **9**, 275–290.

Perruchet, C. (1983) Significance tests for clusters: overview and comments. In *Numerical Taxonomy*. (J. Felsenstein, ed.) pp. 199–208. NATO ASI Series, Vol. G1. Springer-Verlag, Berlin.

Smouse, P. E., Long, J. C. and Sokal, R. R. (1986) Multiple regression and correlation extensions of the Mantel test of matrix correspondence. *Systematic Zoology*, **35**, 627–632.

ter Braak, C. J. F. (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, **67** 1167–1179.

van den Wollenberg, A. L. (1977) Redundancy analysis. An alternative for canonical correlation analysis. *Psychometrika*, **42**, 207–219.