Computational Statistics (2006) 21:183-185
DOI 10.1007/s00180-006-0258-7

# Editorial

When I have been invited as guest editor for this special issue, I realized that this invitation arrived ten years after my first steps in the statistical analysis for *interval-valued* data.

I came in touch with Edwin Diday and other people working in Symbolic Data in 1995; immediately, I was extremely fascinated by the complex-data potential for carrying information. Before long, I started focusing my attention on continuous variables valued as intervals. However, ten years later I feel strongly and even more that in real world data are *as they are*: we do have neither single-valued nor interval-valued variables. Nature of variables depends on our capabilities to measure them. Thanks to the ideas of Edwin Diday and to his revolutionary Symbolic Data Analysis approach, I saw that in parallel to the tangible world there exists another world that cannot be directly measured. I refer to data that are related to the definition of concepts. These data can exist only in terms of interval data; in fact, these definitions are generated to identify a set of statistical units with their natural variability. Working on interval-data statistical analysis, I ran across other cases in which interval data coding can be profitable.

On the above premise, we can distinguish different sources of interval data. Appropriate identification of interval data nature is ticklish problem and it is necessary to adopt consistent analysis treatments. To simplify the description, we can group interval-data sources into three main categories.

Interval data generated by imprecise measurements or repeated measures. This condition refers to all cases in which we are not able to obtain precise measures of the investigated variables. In order to avoid the variability induced by the measurement errors inflates the actual phenomena variability, variables can be coded in terms of interval data, where the lower and upper limit indicate the lowest and the highest registered measure, respectively. Under this condition, interval spreads are extremely narrow with respect the measured values.

A second situation occurs when data refer to concepts having not practically measurable dimensions. A classical example is given by the technical specifications or by the description of species and different product varieties, more generally. Let us think to an animal species or to a kind of wine: related descriptions are necessarily "imprecise" in order to value the variability accountable to. In the same context we can also include interval data resulting from queries to large or huge databases, where the pair minimum and maximum value replaces the mean value.

Last condition refers to intervals associated to couples of variables that are complementary with respect to a given concept. Trade flow data for a given good are defined by the couple of variables: import, export. Analogously, data collected in the customer satisfaction surveys, where expected and perceived satisfaction scores are indicated with respect to a set of given items.

It turns out clear that the different interval data sources share the properties of being described by a pair of ordered values. It is also evident that these different data sources require different treatments according to their specific nature.

This CS special issue intends to present the different aspects in the statistical treatment of interval valued variables as much as possible. Selected papers are characterized with respect their methodological contribution and their application field. There are three domains of interest: factorial analysis, classification and clustering, linear regression.

I received twenty papers; among these, on the basis of the referees reports and on my personal evaluation ten were selected. Unfortunately, for sake of space some interesting papers have been discarded. Reader will notice that papers are not presented according to the first author alphabetic order; they have been classed according to their methodological approach through statistical analysis of interval data.

The paper from Diday and Billiard opens the issue presenting some basic definition and the most basically concept for interval data. Authors present the descriptive statistics for interval data.

Four papers are focused on clustering approaches dealing with interval data. The paper from Chavent *et al.* and the paper from De Carvalho *et al.* present some *dynamic clustering* algorithms. In their paper, D'Urso and Giordani face the problem of the outliers effects in classification. Their attention is focused on a robust classification algorithm. In the clustering framework, the paper from Irpino and Tontodonato proposes the classical two step analysis (factorial analysis + clustering) generalized to interval data. They discuss many aspects related to the distance measures and their consistency in the interval data analysis approach.

Three papers concern to various issues in classification. Duarte-Silva and Brito present an interesting comparison among linear discriminant analysis methods. In this paper are compared the three most important approaches in the interval data coding. Authors highlight very well advantages and drawbacks of the three approaches. Most of their results can be generalized to other analysis fields. In their paper, Ichino and Ishicawa introduce the Cartesian System Model for inter class analysis in pattern classification problems. Empirical proof of the interval-data statistical analysis usefulness is underlined by the paper of Cariou. The author presents a classification tree approach in the electrical consumption profiling.

The last two papers differ from the others because they tackle the interval data analysis problem using approaches derived from the interval arithmetic theory. More specifically, Gioia and Lauro focus their attention on the Principal Component Analysis and propose a new and original approach treating intervals as they are without any coding. Corsaro and Marino propose a complete and critique overview of the interval arithmetic methods for solving equations systems. Their attention is mainly focused on those problems that are more frequent in the statistical analysis. In particular they present a comparison among three interval arithmetic based methods for computing linear regression parameters.

However, it is worth to say that manifested interests through interval-data statistical treatments still do not have equal in real world problems. In actual fact, despite their consistency with the real world phenomena, there is a small number of applications of interval data analysis in facing real world problems. Among twenty papers, I received only two papers focused on practical aspects and only one is in this issue.

Although many topics have been touched, there are many others of interest or potentially interesting for the interval-data statistical treatment. It appears quite clear that classical approaches based on the analytical problem solutions cannot face with the complexity linked to the interval-data. Future challenges will point the attention on the operational research and linear and non-linear programming techniques. Interesting developments in this direction have already manifested in the interval data treatment and in many other statistical fields.

My special thanks are due to more thirty referees I have involved in the revision process. Their comments were accurate and rigorous; they made the job easier form me. I am also indebted with Rosaria Romano for her help in keeping contacts with authors and referees.

*Macerata, April 2006*

*Francesco Palumbo*
*Dipartimento di Istituzioni Economiche e Finanziarie*
*University of Macerata - ITALY*