Available online at www.sciencedirect.com

ScienceDirect

Fuzzy Sets and Systems 281 (2015) 260–271

FUZZY sets and systems

www.elsevier.com/locate/fss

Fuzziness in data analysis: Towards accuracy and robustness

Ana Colubi *, Gil González-Rodríguez

INDUROT, Dept. of Statistics and OR, Campus de Mieres, University of Oviedo, 3600 Mieres, Spain

Received 2 January 2015; accepted 16 May 2015

Available online 3 June 2015

Abstract

The first aim is to emphasize the use of fuzziness in data analysis to capture information that has been traditionally disregarded with a cost in the precision of the conclusions. Fuzziness can be considered in the data analysis process at various stages, but the main target in this paper will be fuzziness in the data. Depending on the nature of the fuzzy data or the aim to which they are handled, different approaches should be applied. We attempt to contribute to the clarification of such a difference while focusing on the so-called ontic approach in contrast to the epistemic approach. The second aim is to underline the need of considering robust methods to reduce the misleading impact of outliers in fuzzy data analysis. We propose trimming as a general and intuitive method to discard outliers. We exemplify this approach with the case of the ontic fuzzy trimmed mean/variance and highlight the differences with the epistemic case. All the discussions and developments are illustrated by means of a case-study concerning the perception of lengths of men and women.

Keywords: Fuzzy methods; Fuzzy data; Fuzziness; Randomness; Statistics; Robust data analysis; Trimming

1. Introduction

Since fuzzy sets are extensively used in Engineering and Science, they started to be exploited also in data analysis (see, e.g., [1–3]). From the initial step of data collection to the final aim regarding, e.g., pattern discovery or the derivation of predictive models, fuzziness has shown to be useful to model a complementary uncertainty or imprecision, traditionally assimilated to randomness or even ignored due to the lack of tools to handle it. Fuzziness may appear indeed at any stage of the data analysis process. Coppi et al. [4] discuss a complete scheme of various ways in which uncertainty, including fuzziness, is present within the information paradigm and argue the need of complementing traditional methods with fuzziness.

Fuzzy methods are being applied to deal with non-fuzzy data, e.g. in fuzzy clustering [5–8] or in handling real distributions [9,10]. On the other hand, standard methods are being tailored to cope with fuzzy data, e.g. in least-squares regression for imprecise input/output [11–15], fuzzy regression [16–18] and other data analytic problems [19–22]. Finally, fuzzy methods have been developed for fuzzy data, e.g. fuzzy/possibilistic clustering of fuzzy data [23,24]. In

* Corresponding author. Tel.: +34 985 458 118.
E-mail address: colubi@uniovi.es (A. Colubi).

http://dx.doi.org/10.1016/j.fss.2015.05.007
0165-0114/

any of these cases, the consideration of fuzziness, as a generalization of its crisp counterpart, allows a finer analysis, which is translated in practice into more realistic or accurate results.

When fuzzy methods appear in combination with a random generation process of the data collection, they are frequently referred to as "fuzzy statistical analysis". The impact of fuzziness in this area has been made patent in various reviews and position papers (see, e.g., [4,9,25–27]). Such papers show indeed the benefits that fuzziness may provide to the statistical analysis, but they basically refer to "ideal statistics", where data are assumed to be clean, i.e. there are not outliers or distributions with heavy tails. One of the main issues that faces real data analysis is the effect that contamination produces in classical statistical measures or methods. This effect is directly inherited by the methods combining statistics and fuzziness and, for this reasons, some alternatives have been explored in particular problems (see, e.g., [28–30]). Within this context the aim of this paper is twofold:

On one hand, we will focus on the special features that entail the presence of fuzziness in experimental or observational data. We will argue that mainstream data analysis problems and approaches can lead to sounder and more sensible results through the consideration of fuzziness. The nature of fuzzy data, and/or the statistical aims, condition the suitability of the analytic approaches (see, e.g., [25,26]). Thus, we attempt to contribute to the literature by exemplifying such analyses while focusing on the so-called ontic approach. This approach considers fuzzy data as whole entities, in contrast to the epistemic approach, which considers fuzzy data as imprecise measurements of precise data.

On the other hand, as a necessary element at this moment, we intent to underline the role of the robust methods to deal with samples of fuzzy data. We will address, in particular, trimming techniques, which are very intuitive and potentially applicable in many general spaces. As an instance, a case-study concerning the computation of the average perception of lengths will be considered at different stages of the paper.

The rest of the paper is structured as follows. In Section 2 the appearance of fuzziness in data collection will be revisited. Section 3 will be devoted to the consideration of some key standard methods for fuzzy data. Section 4 will address robust methods for fuzzy data. Section 5 illustrates the discussed methods with a case-study. The paper concludes with some remarks gathered in Section 6.

2. Fuzziness in the data

Let B be a reference space, fuzziness on B will be modeled through functions $U : B \to [0, 1]$ so that $U(b)$ represents the (possibly partial) membership degree of b. In this paper we will normally consider $B = \mathbb{R}$ and fuzzy sets extending the notion of (closed) intervals, since intervals are often used to represent (crisp) uncertainty in data analysis (see, e.g., [31–34]). Namely, we will usually assume fuzzy sets belonging to $\mathcal{F}_c(\mathbb{R})$, which is the class of functions $U : \mathbb{R} \to [0, 1]$ so that U_α is a nonempty bounded closed interval of $\mathbb{R}$ for all $\alpha \in [0, 1]$, where U_α denotes the α-level set, that is, $U_\alpha = \{x \in \mathbb{R} \mid U(x) \geq \alpha\}$ for all $\alpha \in (0, 1]$ and $U_0 = \mathrm{cl}\{x \in \mathbb{R} \mid U(x) > 0\}$.

The various ways in which fuzziness may affect (statistical) data have always been a topic of concern. It is generally accepted that human perceptions or other inexact measurement mechanisms may produce uncertainty in collected data, which sometimes can be modeled by using fuzzy sets. There is, however, a critical distinction concerning the ontic/epistemic nature of such fuzzy data. In [26], a deep discussion concerning these issues and a detailed review of the basis and results of the epistemic statistical approach are given.

In practice, the difference between both approaches concerns the object of statistical interest. We will consider an example to clarify that. The example is taken from the experiment "Perceptions" which is described in detail in [35] and whose software and data are freely available at http://bellman.ciencias.uniovi.es/SMIRE/perceptions.html.

Summing up, the experiment consisted in asking different people about his/her perception of the relative length of a rule as the gray segment in Fig. 1 w.r.t. a reference rule (in black in Fig. 1). Since it is a difficult task to give an exact numeric forecast, people were allowed to represent their uncertainty through fuzzy sets. Although the software allows the use of more general (convex) fuzzy sets, the participants were suggested to determine only the 0-level and the 1-level set, so that the fuzzy set is completed by linear interpolation. Thus, Fig. 1 represents the perception of a given person who thinks that the relative length of the rule can definitively be any number between approximately 62% and 67% and thinks that there is no way it can be lower than about 54% of greater than 70%.

The experiment was made by showing sequences of segments located at difference positions to different people and gathering different information, although we will focus here on the perception of a fixed length separated by sex (see Figs. 2 and 3). The sample sizes are $n_1 = 49$ for males and $n_2 = 33$ for females.