A Function To Calculate The Median Of A Sample Of Fuzzy Numbers In R

Sinova, Beatriz
Universidad de Oviedo, Departamento de Estadística e I.O. y D.M.
Calvo Sotelo
Oviedo, 33007, Spain
E-mail: sinovabeatriz.uo@uniovi.es

Lubiano, María Asunción
Universidad de Oviedo, Departamento de Estadística e I.O. y D.M.
Calvo Sotelo
Oviedo, 33007, Spain
E-mail: lubiano@uniovi.es

Trutschnig, Wolfgang
European Centre for Soft Computing, Intelligent Data Analysis and Graphical Models
Gonzalo Gutiérrez Quirós
Mieres, 33600, Spain
E-mail: wolfgang.trutschnig@softcomputing.es

1 Introduction

The outcomes of many real-life random experiments in as different fields as Social, Experimental and Biomedical Sciences, present an imprecision prompted by linguistic inaccuracy in value judgements, limited measuring instruments or the focus on data sets in which each datum is given by a set of values, like fluctuations. Furthermore, there is an underlying imprecision in both the first and the third situations which will be formalized in terms of the scale of fuzzy numbers, a rich, intuitive and easy-to-use scale (including real and interval values as special elements) that provides more accuracy than numerical values in a more expressive way than categorical ones.

Random fuzzy numbers (as a particular case of the random fuzzy sets introduced by Puri and Ralescu (1986) under the name of 'fuzzy random variables') are the generalization of the concept of a random variable. They represent the mathematical mechanism to generate data taking values in the scale of fuzzy numbers. Although their probabilistic aspects have been deeply studied, statistical ones are being still analyzed, recalling the lack of linearity dealing with the usual arithmetic with fuzzy values, a universally acceptable total ordering in the space of fuzzy values and realistic 'parametric' families for their distributions.

The most usual central tendency measure to summarize the information given by a random fuzzy set is the Aumann-type expected value (see, for instance, Körner, 2000), defined by means of an L^2-type distance between fuzzy values that plays the same role as the Euclidean distance in the real-valued case. To go into detail, the Aumann-type expected value is the fuzzy set such that minimizes the expectation of the squared distances (in terms of the L^2-type metric) from itself to all the values the random fuzzy set takes on. Though the use of this measure is supported by the Strong Laws of Large Numbers and the fulfilment of the Fréchet's Principle (w.r.t. the previous L^2-type metric), as well as appropriate properties like the equivariance under linear transformations and the sum of random fuzzy numbers, the Aumann-type expected value is, since it is defined using the real-valued mean, very sensitive to the existence of 'extreme' data or data changes.

The more robust behavior of the median in the real-valued case inspires the generalization of this concept for random fuzzy sets. Since there is no universal ranking in this space, it is not possible to define the median as the middle position value. Nevertheless, the concept can be extended using the alternative definition of the median as the value minimizing the mean Euclidean distance w.r.t. variable values. An L^1-metric will be needed for the previous purpose, to be made in Section 3 together with the subsequent analysis of its properties, after recalling some preliminaries on fuzzy numbers and random fuzzy sets in Section 2. In Section 4, an algorithm implemented in R to calculate the median of a sample of fuzzy numbers, included in the latest version of the package SAFD (which contains functions for the basic operations and estimates involving fuzzy numbers) will be described and how it works will be also explained with the help of an illustrative example. Finally, some concluding remarks will be commented in Section 5.

2 Preliminaries on fuzzy numbers and random fuzzy sets

In this Section, only the mathematical definition of the concepts referred to in the Introduction is written because of the restricted paper length. Anyway, some of the features of the space used have already been mentioned.

A *fuzzy number* is a mapping $\widetilde{U}:\mathbb{R}\to[0,1]$ so that for each $\alpha\in(0,1]$ its corresponding α-*level set* $U_\alpha=\{x\in\mathbb{R}:\widetilde{U}\geq\alpha\}$ is a nonempty, closed and bounded interval. One remark is that when it is required, the 0-level is assumed to be $\widetilde{U}_0=\{x\in\mathbb{R}:\widetilde{U}(x)>0\}$. This mapping represents, for each real value x, the 'degree of compatibility of x with the property represented by $\widetilde{U}$' or the 'degree of possibility of the assertion "x is $\widetilde{U}$"'. $\mathcal{F}_c(\mathbb{R})$ will be the notation for the space of fuzzy numbers.

The two more important operations from a statistical point of view, the sum and the product by a scalar, are defined from Zadeh's extension principle (1975). The expressions written below show that the usual fuzzy arithmetic coincides with the level-wise extension of the usual interval-valued operations:

- the *sum* of any two fuzzy numbers $\widetilde{U}$ and $\widetilde{V}$ is defined as the fuzzy number $\widetilde{U}+\widetilde{V}$ such that for each $\alpha\in[0,1]$, $(\widetilde{U}+\widetilde{V})_\alpha=$ Minkowski sum of $\widetilde{U}_\alpha$ and $\widetilde{V}_\alpha=\{y+z:y\in\widetilde{U}_\alpha,z\in\widetilde{V}_\alpha\}$,
- the *product of* $\widetilde{U}$ *by the scalar* γ is defined as the fuzzy number $\gamma\cdot\widetilde{U}$ such that for each $\alpha\in[0,1]$, $(\gamma\cdot\widetilde{U})_\alpha=\gamma\cdot\widetilde{U}_\alpha=\{\gamma\cdot y:y\in\widetilde{U}_\alpha\}$.

As it has already been said, the space $(\mathcal{F}_c(\mathbb{R}),+,\cdot)$ is not linear and to overcome the inexistence of difference (the Hukuhara difference can be considered level-wise, but it may not be well-defined for many fuzzy numbers) distances will be established. Although fuzzy and functional arithmetic are not the same, each element $\widetilde{U}\in\mathcal{F}_c(\mathbb{R})$ can be identified by a function: its support function (see, for instance, González-Rodríguez *et al.*, 2011) $S_{\widetilde{U}}:\{-1,1\}\times(0,1]\to\mathbb{R}$ such that $\widetilde{U}_\alpha=[\inf\widetilde{U}_\alpha,\sup\widetilde{U}_\alpha]=[-s_{\widetilde{U}}(-1,\alpha),s_{\widetilde{U}}(1,\alpha)]$. As a consequence, an L^1-type distance between fuzzy numbers can be defined from the 1-norm between their correspondent support functions, *the 1-norm distance between fuzzy numbers* (δ_1):

$$\begin{array}{rcccl}\delta_1 & : & \mathcal{F}_c(\mathbb{R})\times\mathcal{F}_c(\mathbb{R}) & \to & [0,\infty)\\ & & (\widetilde{U},\widetilde{V}) & \mapsto & \delta_1(\widetilde{U},\widetilde{V}) = \|s_{\widetilde{U}}-s_{\widetilde{V}}\|_1\\ & & & & = \frac{1}{2}\int_{(0,1]}(|\inf\widetilde{U}_\alpha-\inf\widetilde{V}_\alpha|+|\sup\widetilde{U}_\alpha-\sup\widetilde{V}_\alpha|)d\alpha\end{array}$$

which is topologically equivalent to the metric d_1 by Klement *et al.* (1986) defined as follows: $d_1(\widetilde{U},\widetilde{V})=\int_{(0,1]}d_H(\widetilde{U}_\alpha,\widetilde{V}_\alpha)d\alpha$ (where d_H denotes the Hausdorff metric between nonempty compact intervals).

Thus, $(\mathcal{F}_c(\mathbb{R}),\delta_1)$ is a separable metric space. Furthermore, an isometrical embedding of $\mathcal{F}_c^*(\mathbb{R})=\{\widetilde{U}\in\mathcal{F}_c(\mathbb{R}):s_{\widetilde{U}}\in\mathcal{H}_1^*\}$ (where $\mathcal{H}_1^*$ denotes the space of the L^1-type real-valued functions defined on

$\{-1,1\} \times (0,1]$) with the fuzzy arithmetic and the δ_1 metric onto a closed convex cone of $\mathcal{H}_1^*$ with the functional arithmetic and the metric based on the 1-norm is established by means of the mapping $s : \mathcal{F}_c^*(\mathbb{R}) \to \mathcal{H}_1^*$ such that $s(\widetilde{U}) = s_{\widetilde{U}}$. As a consequence, fuzzy data can be treated as functional data by identifying them with their support functions and a lot of results of Functional Data Analysis can be applied if the result belongs to the image of s.

To formalize the process of generation of fuzzy numbers, a *random fuzzy number* (RFN) is defined, given a probability space $(\Omega, \mathcal{A}, P)$, as a mapping $\mathcal{X} : \Omega \to \mathcal{F}_c(\mathbb{R})$ such that, for all $\alpha \in (0,1]$, the α-level mapping $\mathcal{X}_\alpha$ is a compact random interval. Equivalently, it could be defined requiring that $\inf \mathcal{X}_\alpha$ and $\sup \mathcal{X}_\alpha$ are real-valued random variables. Anyway, $\mathcal{X}$ is a Borel-measurable mapping w.r.t. the Borel σ-field generated on $\mathcal{F}_c(\mathbb{R})$ by the topology associated with δ_1 (see Colubi *et al.*, 2001).

As it has already been said, the most usual central tendency measure is the so-called *Aumann-type expected value* of a random fuzzy number: the fuzzy number $\widetilde{E}(\mathcal{X}) \in \mathcal{F}_c(\mathbb{R})$ such that, for all $\alpha \in (0,1]$, $(\widetilde{E}(\mathcal{X}))_\alpha = [E(\inf \mathcal{X}_\alpha), E(\sup \mathcal{X}_\alpha)]$ if these expectations exist.

3 The δ_1 median of a random fuzzy set

To overcome the strong influence data changes and 'extreme' values have on the estimation of the Aumann-type expected value, the generalization of the median will be established as explained in the Introduction, by means of the δ_1 distance. Therefore,

Definition 1 *The median (or medians) of a random fuzzy number $\mathcal{X}$ associated with a probability space $(\Omega, \mathcal{A}, P)$ is (are) the fuzzy number(s) $\widetilde{\mathrm{Me}}(\mathcal{X}) \in \mathcal{F}_c(\mathbb{R})$ satisfying:*

$$E(\delta_1(\mathcal{X}, \widetilde{\mathrm{Me}}(\mathcal{X}))) = \min_{\widetilde{U} \in \mathcal{F}_c(\mathbb{R})} E(\delta_1(\mathcal{X}, \widetilde{U}))$$

The next theorem (see Sinova *et al.*, 2010) proves the existence of at least one median and makes its calculus easier in practice.

Theorem 1 *Given a random fuzzy number $\mathcal{X}$ associated with a probability space $(\Omega, \mathcal{A}, P)$, the fuzzy number $\widetilde{\mathrm{Me}}(\mathcal{X}) \in \mathcal{F}_c(\mathbb{R})$ such that, for all $\alpha \in (0,1]$, $(\widetilde{\mathrm{Me}}(\mathcal{X}))_\alpha = [\mathrm{Me}(\inf \mathcal{X}_\alpha), \mathrm{Me}(\sup \mathcal{X}_\alpha))]$ with the following convention is one median of $\mathcal{X}$ in accordance with the previous definition:*

- *If* $\mathrm{Me}(\inf \mathcal{X}_\alpha)$ *is not unique, it will be chosen to be the midpoint of the interval of medians of* $\inf \mathcal{X}_\alpha$,
- *If* $\mathrm{Me}(\sup \mathcal{X} - \alpha)$ *is not unique, it will be chosen to be the midpoint of the interval of medians of* $\sup \mathcal{X}_\alpha$.

From now, to study its properties, the median will be defined as the unique fuzzy number in the previous theorem. A first remark is that this median needn't coincide with one of the values of the random fuzzy number. An example in which this happens and comments or proofs of the following properties it fulfils can be also found in Sinova *et al.* (2010).

Proposition 1 $\widetilde{\mathrm{Me}}(\gamma \cdot \mathcal{X} + \widetilde{U}) = \gamma \cdot \widetilde{\mathrm{Me}}(\mathcal{X}) + \widetilde{U}$ for all $\gamma \in \mathbb{R}$, $\widetilde{U} \in \mathcal{F}_c(\mathbb{R})$, $\mathcal{X}$ RFN.

That is to say, the median is equivariant by 'linear transformations'. A consequence of the Proposition 1 is that if the distribution of the RFN $\mathcal{X}$ is degenerate at a fuzzy number $\widetilde{U} \in \mathcal{F}_c(\mathbb{R})$ (i.e., $\mathcal{X} = \widetilde{U}$ a.s.[P]), then $\widetilde{\mathrm{Me}}(\mathcal{X}) = \widetilde{U}$.

Although the median couldn't be extended as a 'middle position' value, it is interesting to note that the median just defined can be formalized as a 'middle position' value w.r.t. the *fuzzy max partial order*, stated by Ramík and Římaánek (1985), whenever this order applies:

$\widetilde{U} \precsim \widetilde{V}$ if and only if $\lambda \sup \widetilde{U}_\alpha + (1-\lambda) \inf \widetilde{U}_\alpha \leq \lambda \sup \widetilde{V}_\alpha + (1-\lambda) \inf \widetilde{V}_\alpha$ for all $\alpha, \lambda \in [0,1]$.

Such an order, although partial, is the natural level-wise extension of the product order on $\mathbb{R}^2$ using the inf/sup characterization of the α-levels, so it is relevant to preserve it in the case it applies.

A very important inferential property the median inherits from the real case is its δ_1-strong consistency (if $\mathrm{Me}(\inf \mathcal{X}_\alpha)$ and $\mathrm{Me}(\sup \mathcal{X}_\alpha)$ are unique for every α):

Proposition 2 *Let $\mathcal{X}$ be an RFN associated with a probability space $(\Omega, \mathcal{A}, P)$ such that $\mathrm{Me}(\inf \mathcal{X}_\alpha)$ and $\mathrm{Me}(\sup \mathcal{X}_\alpha)$ are unique for every α (without applying the convention adopted in Theorem 1). Then,*

$$\lim_{n\to\infty} \delta_1\Big(\widehat{\widetilde{\mathrm{Me}}(\mathcal{X})}_n, \widetilde{\mathrm{Me}}(\mathcal{X})\Big) = 0 \quad a.s.\,[P],$$

where $\widehat{\widetilde{\mathrm{Me}}(\mathcal{X})}_n$ denotes the sample median having considered a simple random sample from $\mathcal{X}$.

The main advantage of the sample median of a RFN (as an estimator of the population median) is its higher robustness w.r.t. the sample mean of an RFN (as an estimator of the population mean), shown by their respective finite sample breakdown points (fsbp), which inform of the minimum proportion of sample data that should be perturbed to make the corresponding estimator arbitrarily big or small. Adapting the definition of fsbp given in Donoho and Huber (1983),

Proposition 3 *The finite sample breakdown point of the sample median of an RFN $\mathcal{X}$,*

$$\mathrm{fsbp}(\widehat{\widetilde{\mathrm{Me}}(\mathcal{X})}_n, \widetilde{x}_n, \delta_1) = \frac{1}{n} \min\left\{ k \in \{1,\dots,n\} \,:\, \sup_{Q_{n,k}} \delta_1(\widehat{\widetilde{\mathrm{Me}}(P_n)}, \widehat{\widetilde{\mathrm{Me}}(Q_{n,k})}) = \infty \right\},$$

is $\frac{1}{n} \cdot \lfloor \frac{n+1}{2} \rfloor$ (where $\lfloor\cdot\rfloor$ denotes the floor function, $\widetilde{\mathbf{x}}_n$ denotes the considered sample of n data from the metric space $(\mathcal{F}_c(\mathbb{R}), \delta_1)$ which fulfils that $\sup_{\widetilde{U},\widetilde{V}\in\mathcal{F}_c(\mathbb{R})} \delta_1(\widetilde{U},\widetilde{V}) = \infty$, P_n is the empirical distribution of $\widetilde{\mathbf{x}}_n$ and $Q_{n,k}$ is the empirical distribution of sample $\widetilde{\mathbf{y}}_{n,k}$ obtained from the original one $\widetilde{\mathbf{x}}_n$ by perturbing at most k components).

So taking into account that the finite sample breakdown point of the sample mean of an RFN is $\mathrm{fsbp}(\overline{\mathcal{X}_n}, \widetilde{x}_n, \delta_1) = \frac{1}{n}$, the finite sample breakdown point of the sample mean from an RFN $\mathcal{X}$ is lower than the one for the sample median when the sample size is $n > 2$.

4 An R function to calculate the median of a sample of fuzzy numbers

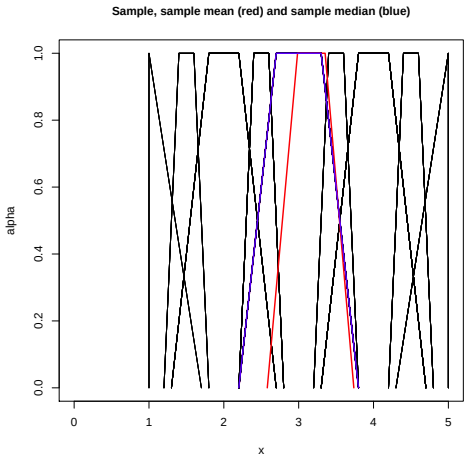
In this section, the functionality of the R function `Fmedian` to compute the median of a given sample of fuzzy numbers is explained: the more important parts of its R algorithm, contained in the package SAFD (see Trutschnig and Lubiano, 2011) will be commented and some illustrative examples will show how it may be used.

The median is the 50%-quantile, so the algorithm `Fmedian` calls the function `Fquantile`, also contained in the SAFD package. Specifying the data set and the quantiles one wants to compute, `Fquantile` first computes the (Minkowski) mean if the data are in the correct form (in case not, the function `checking` informs of whether the fuzzy numbers introduced are in fact fuzzy numbers and `translator` writes all of them using the same number of α-levels).

Sample data are collected in a matrix in which columns represent the sorted x-values of the infima and the suprema of the considered α-levels for all sample fuzzy data. Quantiles will be computed levelwise, i.e., for each row in the matrix.

As an illustrative example the data set `Trees` contained in SAFD can be loaded. It is a list of three sublists which inform about the quality of the three main species of trees in Asturias (birch, sessile oak and rowan) in a reforested area. Let's consider the first sublist: a sample of 133 trapezoidal

fuzzy numbers. Each of them represent the experts subjective perception of the tree quality on a scale from 0 to 5 (0 associated with very bad quality and 5, with very good quality). Recall that the 1-cut is the interval in which the expert believes the quality of the tree to be contained in and the 0-cut, the interval in which he/she is totally sure the quality is contained. The median is obtained and plotted (together with the Aumann-type expected value):



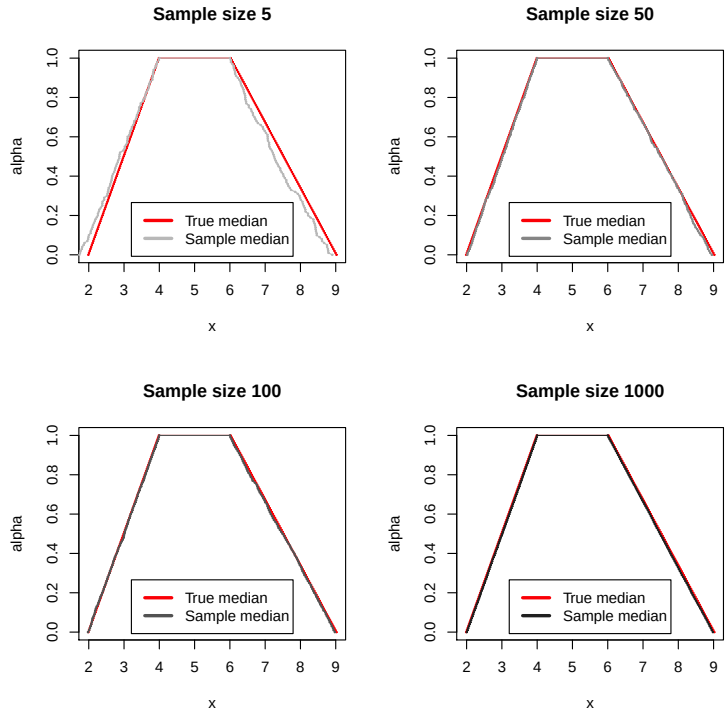

```
data(Trees)
Sample<-Trees$species1
Fmedian(Sample,pic=1)
```

	x	alpha
1	2.20	0.00
2	2.70	1.00
3	3.30	1.00
4	3.80	0.00

The second example shows how the sample median converges to the population one when the sample size increases. In the code it can be seen how the `generator` function (from SAFD package) is used to generate a sample of a population with a known mean, `X`, and with a population median easily determined because of the properties of the gamma distribution and the special choice of `X`:

```
XL<-data.frame(cbind(x=c(2,4,4,4),alpha=c(0,1,1,0)))
XR<-data.frame(cbind(x=c(6,6,6,9),alpha=c(0,1,1,0)))
nl<-101
EL<-translator(XL,nl)
ER<-translator(XR,nl)
R<-c(5,50,100,1000)
A<-list()
SS<-list()
for (k in 1:length(R)){
 SSL<-vector("list",length=R[k])
 SSR<-vector("list",length=R[k])
 for (j in 1:R[k]){
SSL[[j]]<-generator(EL,pertV=list(dist="unif",par=c(0,0)),pertL=list(dist="exp",par=c(1)),
pertR=list(dist="exp",par=c(1)))
SSR[[j]]<-generator(ER,pertV=list(dist="unif",par=c(0,0)),pertL=list(dist="exp",par=c(1)),
pertR=list(dist="exp",par=c(1)))
SS[[j]]<-data.frame(cbind(x=c(SSL[[j]]$x[1:nl],SSR[[j]]$x[(nl+1):(2*nl)]),alpha=EL$alpha))}
M<-Mmean(SS,pic=0)
A[[k]]<-Fmedian(SS,pic=0)
A[[k]]}
quants<-rep(0,(2*nl))
for(i in 1:(nl)){quants[i]<-mean(XL$x[2:3])-qgamma(0.5,shape=i,scale=2/(nl-1))}
for (i in (nl+1):(2*nl)){quants[i]<-mean(XR$x[2:3])+qgamma(0.5,shape=i-nl,scale=3/(nl-1))}
xq<-sort(quants[1:(nl)])
quants<-c(xq,quants[(nl+1):(2*nl)])
Q<-data.frame(x=quants,alpha=A[[1]]$alpha)
dev.new()
par(mfrow=c(2,2))
for (k in 1:length(R)){
 plot(Q,type="l",cex=0.1, col="red",main=paste("Sample size ",R[k],sep=""),xlim=c(2,9))
 lines(Q,type="p",col="red",cex=0.2)
 lines(A[[k]],type="l",cex=0.1,col=colors()[225-20*(k-1)])
```

```
lines(A[[k]],type="p",col=colors()[225-20*(k-1)],cex=0.2)
 legend("bottom", c("True median", "Sample median"), inset = c(.03, .03), lwd=3, lty=1,
col=c('red',colors()[225-20*(k-1)]),)
```



References

[1] Colubi, A., Domínguez-Menchero, J.S., López-Díaz, M., Ralescu, D.A., 2001. On the formalization of fuzzy random variables. Inform. Sci. 133, 3–6.

[2] Donoho, D.L., Huber, P.J., 1983. The notion of breakdown point. In: A Festschrift for Erich L. Lehmann (Bickel, P.J., Doksum, K., Hodges, J.L. Jr. eds.). Wadsworth, Belmont, 157–184.

[3] González-Rodríguez, G., Colubi, A., Gil, M.A., 2011. Fuzzy data treated as functional data. A one-way ANOVA test approach. Comput. Stat. Data Anal. In press (doi:10.1016/j.csda.2010.06.013).

[4] Klement, E. P., Puri, M. L., Ralescu, D. A., 1986. Limit theorems for fuzzy random variables. Proc. R. Soc. Lond. A 407, 171-182.

[5] Körner, R., 2000. An asymptotic α-test for the expectation of random fuzzy variables. J. Stat. Plann. Inference 83, 331–346.

[6] Puri, M.L., Ralescu, D.A. 1986. Fuzzy random variables. J. Math. Anal. Appl. 114, 409–422.

[7] Ramík, J., Římánek, J., 1985. Inequality relation between fuzzy numbers and its use in fuzzy optimization. Fuzzy Sets and Systems 16, 123-138.

[8] Sinova, B., Gil, M.A., Colubi, A., González-Rodríguez, G., Van Aelst, S., 2010. An approach to the median of a random fuzzy set. Abstracts of 3rd International Conference of the ERCIM Working Group on Computing & Statistics, p.97. (`http://www.cfe-csda.org/cfe10/LondonBoA.pdf`)

[9] Trutschnig, W., Lubiano, M.A., 2011. SAFD: Statistical Analysis of Fuzzy Data (R package) (`http://cran.r-project.org/web/packages/SAFD/index.html`).

[10] Zadeh, L.A., 1975. The concept of a linguistic variable and its application to approximate reasoning, Part 1. Inform. Sci. 8, 199–249; Part 2. Inform. Sci. 8, 301–353; Part 3. Inform. Sci. 9, 43–80.