STUDIA UBB CHEMIA, LXV, 2, 2020 (p. 45-56)
(RECOMMENDED CITATION)
DOI:10.24193/subbchem.2020.2.04

MINERAL WATERS CLASSIFICATION USING FUZZY LINEAR DISCRIMINANT ANALYSIS

ALEXANDRINA GUIDEA[a], RADU D. GĂCEANU[b], HORIA F. POP[b], COSTEL SÂRBU[a]*

ABSTRACT. Fuzzy linear discriminant analysis is efficiently applied for the characterization and classification of some Romanian and German mineral waters according to their mineral composition. The samples were successfully classified according to the degrees of membership and canonical scores. A correct classification rate of 88% was obtained when the samples were divided into four groups corresponding to origin and nature of samples. The proposed methodology based on the fuzzy sets theory may be considered as a promising tool with future applications in analytical chemistry and other related fields.

Keywords: *Fuzzy discriminant analysis, chemometrics, mineral waters, mineral composition*

INTRODUCTION

One of the most acute problems facing the world today is the water quality and quantity, because water is a limiting factor of the environment, both for biological systems and human societies [1-3]. As a consequence, there are many national and international initiatives and vigorous efforts to protect ground and surface water and to increase water quality and resources [4]. Mineral waters, as natural waters usually obtained from springs, contain an appreciable quantity of salts and gases deriving from their passage through rocks and soil. From the physiological point of view, mineral waters

[a] *Babeş-Bolyai University, Faculty of Chemistry and Chemical Engineering, 11 Arany Janos str., RO-400028, Cluj-Napoca, Romania*
[b] *Babeș-Bolyai University, Department of Computer Science, str. Mihail Kogalniceanu nr. 1, 400084, Cluj-Napoca, Romania*
** Corresponding author: csarbu@chem.ubbcluj.r*

must contain a sufficient amount of inorganic salts, with or without dissolved gases, to enable them to have an efficient effect [5]. There are many mineral waters, presently being bottled all around the World. The most famous in Europe, for example, are from France, Italy, Germany, Austria and Switzerland. Compared to Europe's mineral waters, Romanian natural mineral waters take a very high position by their diversity, quality and quantity. Romanian mineral waters are a gift of untouched nature in close surroundings of the Carpathian Mountains from North to South [6, 7]. Tradition of utilization of these waters for drinking originates from the Roman period in Dacia and is connected for practical comprehension of its healing activity on the work of heart and digestive organs, for example. In the past, many of these waters were neither present on the market, nor subject of any substantial examinations. It is interesting that despite the existence of a general chemical analysis (macro components) of the majority of these waters it has not been noticed that many of them contain increased contents of some micro elements as, for example, magnesium, zinc, copper or fluoride and bromine. Relatively recent, research has shown that the status of these elements in the human organism in its development phases has extremely great importance in prevention of numerous illnesses. For example, the prominent role of Mg in water as a cardiovascular protective factor is largely accepted [8].

The classical discriminant analysis method is known to provide maximum likelihood estimations under certain assumptions (normality of the class distributions etc.) [9-11]. As the experiments will illustrate, and as previous research on data analysis methods based on fuzzy sets have also shown, the fuzzy discriminant analysis method is robust with respect to outliers and distribution of data [12, 13].

We underline once again the robustness achieved by using fuzzy membership values and their relevance. The main advantage of fuzzy sets over crisp sets and of fuzzy logic over binary logic is the availability of gradual membership degrees. On one side, the classes input provided by the human expert is fuzzified, allowing robust treatment of outliers. On the other side, the output of the method is fuzzy as well, allowing for a more detailed view of the relationships between data objects (samples) and classes. These fuzzy degrees of membership (DOMs) are not actually related to uncertainty, because there is nothing uncertain about the classification of a certain data element (sample), but have to be regarded as a measure of 'typicality' [12-14].

The fuzzy linear discriminant analysis (FLDA) method presented and applied here is a multiclass method by design, as no restrictions with respect to the number of classes are introduced. This is a parameter to be set by the human experts as they establish the a-priori classes split.

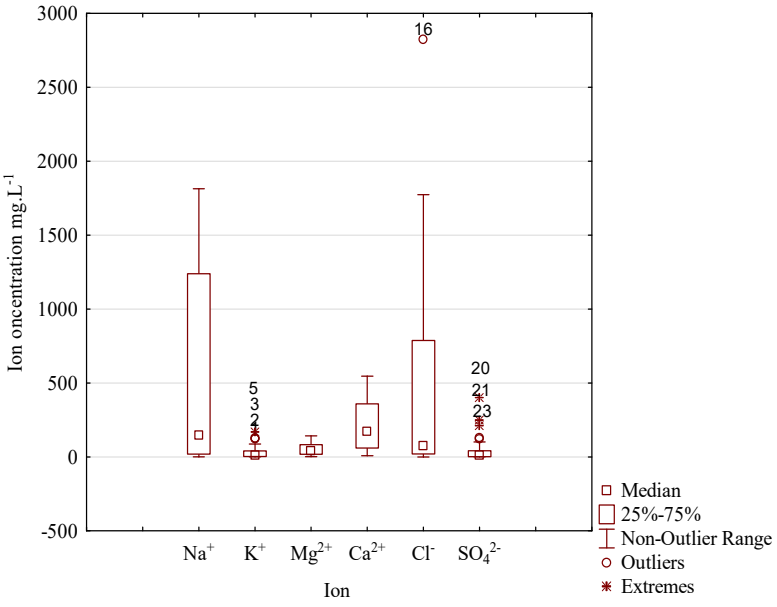
RESULTS AND DISCUSSION

The data used in this study are summarized in **Table 1** and **Figure 1a-b.** It can be easily observed from the Box and whiskers plots that only in the case of major ions (Na^+, Ca^{2+}) including Mg^{2+}, Sr^{2+} and Cu^{2+} and also conductivity and salinity are not highlighted outliers and extremes values (Fig. 1a-b). In all these cases, all values are within the robust confidence interval (neither outliers nor extremes) and the distributions appear to be more or less asymmetric. In all other cases are highlighted outliers and extremes values. The concentrations of Li^+, K^+, NH_4^+, Sr^{2+}, Ba^{2+}, Cu^{2+} and Br^- are the highest in samples from Sângeorz Băi area and the concentration of SO_4^{2-} is the highest in mineral waters from Germany.

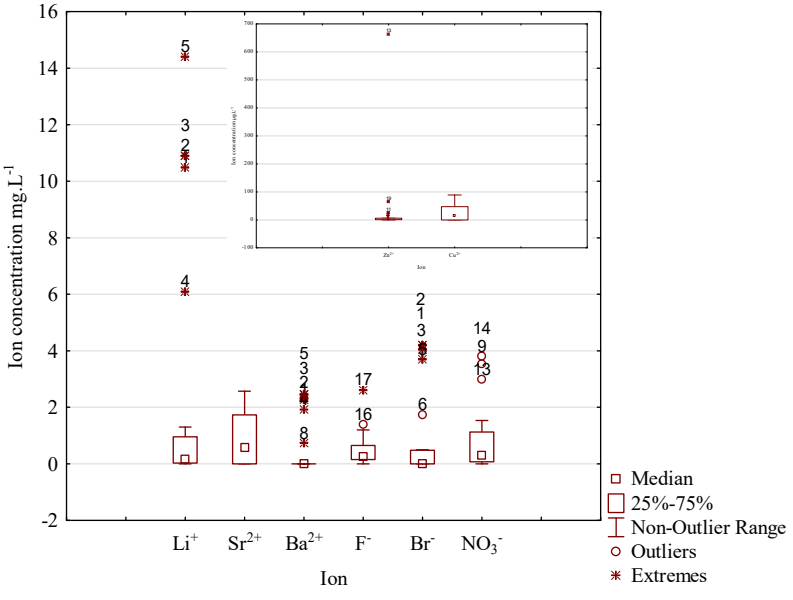
Table 1. The statistics of data corresponding to all mineral waters discussed in this study

Variable	Mean	Median	Minimum	Maximum	Range	SD	Skewness	Kurtosis
Li^+	2.3	0.2	0.0	14.4	14.4	4.4	1.8	2.0
Na^+	514.4	145.0	0.9	1814.0	1813.1	696.2	1.1	-0.6
K^+	35.0	13.0	0.5	166.0	165.5	49.2	1.6	1.2
NH_4^+	3.0	0.1	0.0	18.0	18.0	5.1	1.7	2.2
Mg^{2+}	52.2	42.5	1.7	142.0	140.3	41.0	0.5	-0.9
Ca^{2+}	214.4	171.0	8.8	546.0	537.2	172.7	0.4	-1.3
Sr^{2+}	0.8	0.6	0.0	2.6	2.6	0.8	0.8	-0.8
Ba^{2+}	0.5	0.0	0.0	2.5	2.5	0.9	1.6	0.6
Zn^{2+}	34.0	3.5	0.0	664.0	664.0	132.0	4.9	24.4
Cu^{2+}	24.9	16.0	0.0	89.0	89.0	29.4	0.8	-0.8
Cl^-	514.6	76.3	0.2	2824.0	2823.8	819.5	1.5	1.3
F^-	0.5	0.3	0.0	2.6	2.6	0.6	2.4	6.8
Br^-	0.8	0.0	0.0	4.2	4.2	1.5	1.8	1.6
SO_4^{2-}	56.9	9.8	0.6	398.0	397.4	97.8	2.4	5.7
NO_3^-	0.8	0.3	0.0	3.8	3.8	1.1	1.9	2.7
pH	6.8	6.9	5.2	8.4	3.3	0.9	0.0	-1.0
Cond	3407.4	1802.0	104.0	10680.0	10576.0	3560.5	1.1	-0.4
Salinity	1.8	0.9	0.0	6.0	6.0	2.0	1.1	-0.4

All concentrations are in $mg.L^{-1}$, excepting Cu^{2+} and Zn^{2+} which are in $\mu g.L^{-1}$



(**a**)



(**b**)

Figure 1. Box and whiskers plot of the physicochemical parameters (**a**, **b**) excepting sample S1 assigned also to German mineral water group (**A4**) but with a very small DOM (0.2654). This sample appears as a strong outlier with an equal DOM to other fuzzy partitions. All the German mineral water samples were assigned to fuzzy partition **A4** with DOM-range between 0.8959 and 0.9970, excepting in this case the sample G2 (**21**-Schiller Brunnen) assigned to the Olăneşti group with a high DOM (0.8782).

According to the origin and nature of mineral water samples, the number of classes for FLDA was chosen to be 4. FLDA produced four fuzzy partitions, which were all represented by a prototype (a cluster center with the parameters corresponding to the fuzzy robust means of the original physicochemical characteristics for the 25 samples weighted by DOMs corresponding to each partition) depicted in **Table 2**. To compare the partitions and the similarity and differences of the investigated mineral waters, we have to analyze both the characteristics of the prototypes corresponding to the four fuzzy partitions (**A1**-**A4**) obtained by applying FLDA and DOMs of samples corresponding to all fuzzy partitions, including also the canonical scores used usually in classical linear discriminant analysis. The results presented in Table 2 clearly illustrate the most specific characteristics of each fuzzy partition and their similarity and differences. The values of prototype corresponding to the first partition (**A1**) assigned to medicinal waters from Sângeorz Băi are the highest excepting the value of Zn^{2+}, F^- and SO_4^{2-}. The highest value of Zn^{2+} corresponds to the prototype of **A2** partition (assigned to the table Romanian waters-M), the highest value of F^- to the prototype of **A3**

Table 2. The parameters of prototypes

Physicochemical parameter	Parameters of prototype			
	A1	A2	A3	A4
Li^+	10.38	0.35	0.06	0.23
Na^+	1720.84	91.60	250.41	33.98
K^+	124.56	12.44	5.07	7.12
NH_4^+	10.56	1.08	0.72	0.09
Mg^{2+}	90.55	42.02	28.31	31.96
Ca^{2+}	435.74	187.02	47.98	126.67
Sr^{2+}	1.69	0.61	0.00	0.66
Ba^{2+}	2.19	0.07	0.00	0.00
Zn^{2+}	1.16	83.12	33.60	4.16
Cu^{2+}	57.71	22.58	0.00	15.93
Cl^-	1564.11	47.68	409.55	31.18
F^-	0.66	0.28	0.98	0.17
Br^-	3.57	0.16	0.00	0.00
SO_4^{2-}	1.89	10.60	107.34	142.10
NO_3^-	0.65	1.67	0.03	0.16
pH	6.98	6.28	7.78	6.89
Cond	9544.16	1479.77	1713.96	852.26
Salinity	5.35	0.73	0.81	0.37

(assigned to the mineral waters from Olăneşti-O area) and the highest concentrations of SO_4^{2-} characterize the German mineral waters-G assigned to the **A4** partition.

The results presented in **Table 3** and **Figure 2a-c** point out the mineral waters assigned to each partition (**A1**-**A4**) according to their DOMs.

The fuzzy partition **A1** includes all the *samples* from Sângeorz Băi area with the DOM-range between 0.9411 and 0.9611 for the first five samples and 0.7705 for the **6**-Anies sample. This is not surprisingly because Anies is located at seven kilometers far from Sângeorz Băi. The fuzzy partition **A2** contains eight table Romanian waters (M1, M8) with the DOM-range between 0.8975 and 0.9971 excepting the sample M9 assigned to **A4** (German mineral water group) but with a relatively small DOM (0.5570). The fuzzy partition **A3** includes three samples from Olăneşti area (S2, S3 and S4) with the DOM-range between 0.8439 and 0.9594,

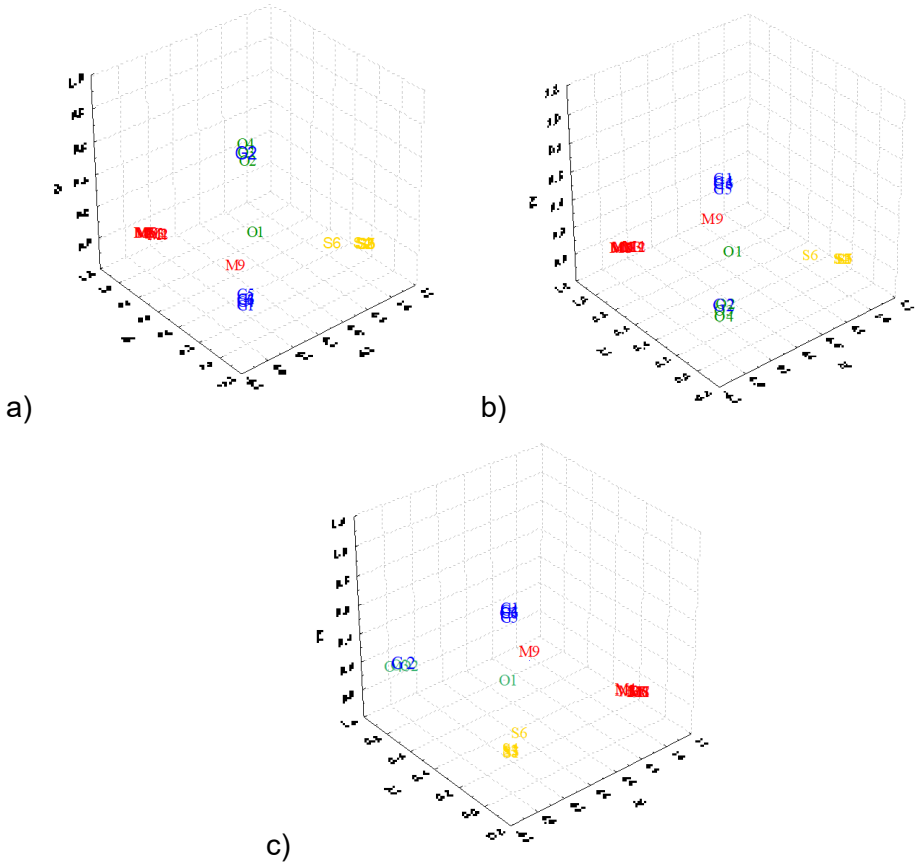


Figure 2. 3-D scatterplot of DOMs

Table 3. The degrees of membership (DOMs) to the four fuzzy classes obtained applying FLDA

No	Water sample	**A1**	**A2**	**A3**	**A4**
1	**S1**	**0.9695**	0.0111	0.0075	0.0119
2	**S2**	**0.9540**	0.0125	0.0142	0.0193
3	**S3**	**0.9659**	0.0093	0.0103	0.0145
4	**S4**	**0.9411**	0.0153	0.0186	0.0250
5	**S5**	**0.9711**	0.0104	0.0071	0.0114
6	**S6**	**0.7705**	0.1066	0.0451	0.0778
7	**M1**	0.0010	**0.9971**	0.0005	0.0014
8	**M2**	0.0380	**0.8992**	0.0203	0.0426
9	**M3**	0.0291	**0.9242**	0.0149	0.0318
10	**M4**	0.0313	**0.8975**	0.0184	0.0528
11	**M5**	0.0081	**0.9775**	0.0044	0.0100
12	**M6**	0.0050	**0.9865**	0.0026	0.0060
13	**M7**	0.0027	**0.9925**	0.0014	0.0034
14	**M8**	0.0025	**0.9928**	0.0014	0.0033
15	M9	0.1101	0.2332	0.0997	0.5570
16	O1	*0.2584*	*0.2242*	*0.2520*	*0.2654*
17	**O2**	0.0354	0.0224	**0.8439**	0.0982
18	**O3**	0.0135	0.0094	**0.9080**	0.0691
19	**O4**	0.0064	0.0039	**0.9594**	0.0303
20	**G1**	0.0006	0.0007	0.0017	**0.9970**
21	G2	0.0141	0.0100	***0.8782***	*0.0977*
22	**G3**	0.0116	0.0147	0.0240	**0.9497**
23	**G4**	0.0064	0.0075	0.0143	**0.9718**
24	**G5**	0.0265	0.0343	0.0433	**0.8959**
25	**G6**	0.0143	0.0169	0.0269	**0.9419**

All of the above statements concerning the efficiency of FLDA are well supported also by scatterplot of canonical scores in the space defined by Froot 1 - Froot 2 - Froot 3 **(**Figure 3) and the values of quality performance features obtained for the correct classification rate of the original data and by applying the leave-one-out (LOO) cross-validation approach (**Table 4**).

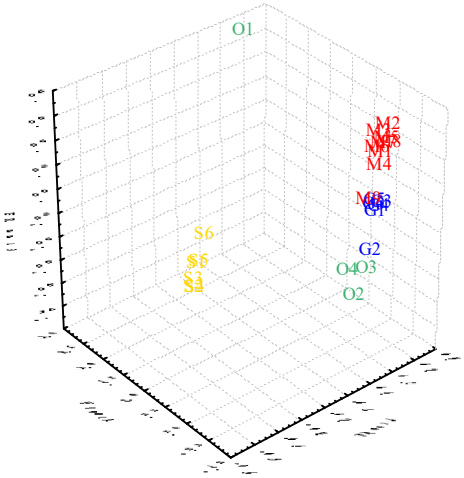


Figure 3. 3-D scatterplot of canonical scores obtained applying FLDA

Table 4. Matrix classification of mineral waters

Class	Total	Samples				%			
		A1	A2	A3	A4	A1	A2	A3	A4
A1	**6**	**6**	0	0	0	**100.00**	0.00	0.00	0.00
A2	**9**	0	**8**	0	1	0.00	**88.89**	0.00	11.11
A3	**4**	0	0	**3**	1	0.00	0.00	**75.00**	25.00
A4	**6**	0	0	1	**5**	0.00	0.00	16.67	**83.33**
Leave-one-out cross-validation									
A1	**6**	**3**	2	1	0	**50.00**	33.33	16.67	0.00
A2	**9**	0	**5**	2	2	0.00	**55.56**	22.22	22.22
A3	**4**	0	0	**1**	3	0.00	0.00	**25.00**	75.00
A4	**6**	0	0	2	**4**	0.00	0.00	33.33	**66.67**

CONCLUSION

In this study, the advantages of the fuzzy linear discriminant analysis for the characterization and classification of various mineral waters on the basis of their mineral composition have been explored. The parameters of the prototypes (class centers) illustrate much better than, for example, arithmetic means the specific characteristics of each class, and the degrees of membership allow a rationale comparison of the similarities and differences of mineral water samples investigated.

EXPERIMENTAL SECTION

Mineral water samples and analytical methods

A large diversity of natural mineral waters from Romania (19 types) were analyzed and compared (**1**-**19**): **1**-Sângeorz Băi spring 1 (S1), **2**-spring 3 (S2), **3**-spring 5 (S3), **4**-spring 7 (S4), **5**-spring 9 (S5) and **6**-Anies (S6); **7**-Olăneşti spring 30 (O1), **8**-spring 14 (O2), **9**-spring 24 (O3) and **10**-spring 10 (O4); **11**-Borsec (M1), **12**-Biborţeni (M2), **13**-Bucovina (M3), **14**-Anavie (M4), **15**-Dorna (M5), **16**-Poiana Negri (M6), **17**-Buziaş (M7), **18**-Izvorul Alb (M8) and **19**-Izvorul Minunilor (M9).

Only six types of German mineral waters (**20**-**25**) were included in this study: **20**-Fuldataler (G1), **21**-Schiller Brunnen (G2), **22**-Lauchaer Minnerall Brunnen (G3), **23**-Schönborn Quelle (G4), **24**-Lausitzer (G5) and **25**-Schildetaler Mineralquelle (G6).

It is needed to mention that all Romanian M type waters are bottled nowadays on the market and are very appreciated by the public people. The waters from Sângeorz Băi and Olăneşti region respectively, well-known as natural medicinal waters, are recommended especially in the cure of gastroenterolitic disorder, cholecysitis, hyperuricemia, consequences of liver disease, disorder of biliar tract and gastroenterolitic disorder.

All samples were analyzed for the anions, chloride, fluoride, bromine, sulfate, nitrate and cation lithium, sodium, potassium, ammonium, magnesium, calcium, strontium and barium by ion chromatography using a DIONEX DX 120 system (Methrom anion and cation columns, deionized water and acetonitrile as eluent and a conductivity detector). The concentration of cooper was determined by standard flame atomic absorption spectrometry (Perkin Elmer FIAS 400), using specific line and the concentration of zinc by standard striping voltammetry (Metrohm Polarecord 626); pH and conductivity were determined electrochemically [9, 15]. The results obtained for the samples presented above are depicted in Tables 1 and 2.

Fuzzy Linear Discriminant Analysis

The Fuzzy Linear Discriminant Analysis problem is defined as follows: let $\mathbf{X} = \{x^1, ..., x^n\} \subset \mathbf{R}^s$ be a finite set of characteristic vectors, where n is the number of items and s is the number of the original variables (predictors), $\mathbf{x}^j = [x^j_1, x^j_2, ..., x^j_s]^T$ and let A_i (with $i = 1, ..., k$) be fuzzy sets on X, corresponding to the k a-priori sets composing the partition substructure of the given data set. A new vector (or characteristic) c is to be determined, that maximizes the fuzzy between-class variance of the projected data items, and minimizes the fuzzy within-class variance of the projected data items.

The total variance/covariance matrix may be decomposed into two components: the between-group variance **B** and within-group variance **W**, namely,

$$\mathbf{V} = \mathbf{B} + \mathbf{W}.$$

Considering a new characteristic defined as **c** = **Xu**, this becomes

$$\frac{\mathbf{u}^T\mathbf{B}\mathbf{u}}{\mathbf{u}^T\mathbf{V}\mathbf{u}} + \frac{\mathbf{u}^T\mathbf{W}\mathbf{u}}{\mathbf{u}^T\mathbf{V}\mathbf{u}} = 1$$

With the first ratio maximized we get

$$\lambda = \frac{\mathbf{u}^T\mathbf{B}\mathbf{u}}{\mathbf{u}^T\mathbf{V}\mathbf{u}}$$

Or, since matrix **V** of the total variance is symmetrical and positive definite,

$$\mathbf{V}^{-1}\mathbf{B}\mathbf{u} = \lambda\mathbf{u},$$

where λ ($0 \le \lambda < 1$) and **u** represent the eigenvalues (known, as well, as characteristic roots) and eigenvectors of the matrix $\mathbf{V}^{-1}\mathbf{B}$.

The vector $\mathbf{u}^1$, named *the first discriminant factor* corresponds to the highest value of λ; the higher this value the higher will be the discriminant power of this factor.

Considering this new characteristic defined as **c** = **Xu**, the fuzzy between-group variance **B** and fuzzy within-group variance **W**, are defined as:

$$\mathbf{W} = \frac{1}{n-k}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} A_i\left(x^j\right)^m\left(x^j - L^i\right)^T\left(x^j - L^i\right)\right),$$

$$\mathbf{B} = \frac{1}{k-1}\sum_{i=1}^{k}\left(\sum_{j=1}^{n} A_i\left(x^j\right)\right)^m\left(L^i - L\right)^T\left(L^i - L\right),$$

where the class means L^i are the fuzzy central locations of classes A_i, and L is the central location for the whole data set. The weighting exponent m (so-called "fuzzifier") is any real number in [1, ∞], which determines the fuzziness of the clusters (for $m \to 1$ the μ_{ij} approach 0 or 1, for $m \to \infty$ the memberships tend to be "totally fuzzy" $\mu_{ij} \to 1/c$). Usually, m = 2.

As the fuzzy sets A_i form a sub-partition of the given data set, we formulate the problem of determining the optimal direction **u** as maximizing the ratio

$$\lambda = \frac{\mathbf{u}^T(\mathbf{V}-\mathbf{W})\mathbf{u}}{\mathbf{u}^T\mathbf{V}\mathbf{u}}$$

with ($0 \le \lambda < 1$). In a different form, since matrix **V** of the total variance is symmetrical and positive definite,

$$\mathbf{V}^{-1}\mathbf{(V-W)u} = \lambda\mathbf{u},$$

where λ and **u** represent the eigenvalues (known, as well, as characteristic roots) and eigenvectors of the matrix $\mathbf{V}^{-1}\mathbf{(V-W)}$.

The vector $\mathbf{u}^1$, named *the first fuzzy discriminant factor* corresponds to the highest value of λ; the higher this value the higher will be the discriminant power of this factor. After obtaining the first discriminant characteristic $\mathbf{c}_1 = \mathbf{Xu}^1$, in a similar way can be obtained the discriminant characteristic $\mathbf{c}_2 = \mathbf{Xu}^2$, uncorrelated with the first and so on. It appears clearly that eigenvectors corresponding to the matrix $\mathbf{V}^{-1}\mathbf{(V-W)}$ namely $\mathbf{u}^1$, $\mathbf{u}^2$, .., $\mathbf{u}^{k-1}$, ranked in decreasing order of the positive values $\lambda_1, \ldots, \lambda_{k-1}$, are successive solutions of the above matrix equation. The quality of discrimination and the selection of the most discriminant independent variable is given by the value of the largest eigenvalue, λ.

Finally, the original class means are projected in the new system of coordinates, and the final fuzzy membership degrees are determined from square-distances to the class means, as with the Fuzzy *c*-Means algorithm.

The final fuzzy classification table is then computed by counting cardinals of fuzzy sets: instead of counting the number of data items classified in a particular class, we are actually computing an overall fuzzy membership degree. The fuzzy count of all items from the *i*-th original fuzzy set A'_i classified in the *l*-th fuzzy set A_l denoted as C_{il}, is given by

$$C_{il} = \sum_{j=1}^{n} A'_i\left(x^j\right) \cdot A_l\left(x^j\right).$$

or, by scaling the values above as the percentages of all items in A'_i classified in A_l :

$$C_{il}^{[\%]} = \frac{\sum_{j=1}^{n} A'_i\left(x^j\right) \cdot A_l\left(x^j\right)}{\sum_{j=1}^{n} A'_i\left(x^j\right)} \times 100.$$

A crisp classification matrix is as well determined by first defuzzifying the final fuzzy partition and then using the cardinals of the crisp classes. After this learning phase, testing follows in various ways, including use of separate testing data, or by cross-validation.

All the graphs and some statistics were performed using Statistica 8.0 (StatSoft, Inc. 1984–2007, Tulsa, USA) software. All the other results were obtained using our own fuzzy software package.

REFERENCES

1. J.H. Lehr; J. Keeley; Water Encyclopedia - Ground Water, Wiley-Interscience, New York, **2005**.
2. S. Ahuja; Handbook of Water Purity and Quality, Academic Press, **2009**.
3. S. Maxwell; S. Yates; The Future of Water: A Startling Look Ahead, American Water Works Association, **2011**.
4. C.E. Boyd; Water Quality. An Introduction, Second Edition, Springer, **2015**.
5. I. Rosborg; F. Kozisek; (Eds). Drinking Water Minerals and Mineral Balance. Importance, Health Significance, Safety Precautions, Second Edition, Springer Nature Switzerland AG, **2019**.
6. C. Berca; Water and Health, Ceres Publishing House, Bucharest, **1994**.
7. L. Munteanu; C. Stoicescu; L. Grigore; The Guide of Romanian Resting Stations, Sport Turism Publishing House, Bucharest, **1978**.
8. J. Durlach; Magnesium in Clinical Practice, Publ. John Libbey, London-Paris, **1988**.
9. C. Sârbu; *Rev. Chim*. (Bucharest), **2002**, *53*, 442-449.
10. C. Sârbu; H.F. Pop; R. Elekes; G. Covaci; *Rev. Chim*. (Bucharest), **2008**, *59*, 1237-1241.
11. G.J. McLachlan; Discriminant Analysis and Statistical Pattern Recognition, John Wiley & Sons, Inc., **2004**.
12. C. Sârbu; H.F. Pop; Fuzzy Soft-Computing Methods and Their Applications in Chemistry in Reviews in Computational Chemistry, K.B. Lipkowitz, R. Larter and T.R. Cundari (Eds.), Wiley-VCH, 2004, Chapt. 5, 249-332.
13. H.F. Pop; C. Sârbu; A New Fuzzy Discriminant Analysis Method, *MATCH Commun. Math. Comput. Chem*., **2013,** *69*, 391-412.
14. O. Horovitz; C. Sârbu; H.F. Pop; Clasificarea raţională a elementelor chimice, Editura Dacia, Cluj-Napoca, **2000**.
15. C. Sârbu; H.W. Zwanziger; *Anal. Lett*, **2001**, *34*, 1541--1552.