Comparison criteria strengthened in classification and type representation, according to the Statistical Model of Impact Measuring, in a case study in Pastaza, Ecuador

E.O. Segura[1] and Verena Torres[2]

[1]*Universidad Estatal Amazónica, km 2½, Napo, Puyo, Pastaza, Ecuador*
[2]*Instituto de Ciencia Animal, Apartado Postal 24, San José de las Lajas, Mayabeque, Cuba*
Email: edisonsegurachavez@gmail.com

The classifications of five, four, and three groups, obtained using the Statistical Model of Impact Measuring, with the Eta coefficient, were compared, and those corresponding to the confirmatory method of K-Means, which uses the statistical test of Fisher, depending on the estimated impacts of a study of 90 dairy farms. Variable selected were: slope (%), farm area (ha), area used by cattle (ha), compatible area of grazing (%), number of cows (head), female cows for reproduction (head), cows for dairy (head), milk production per year (thousands of liters, mL), number of gullies ha^{-1} and soil depth (cm). The study was developed in Pastaza province, Ecuador. It was confirmed that for the classifications in five and four groups, 90% of Eta values were higher than 0.60, and for the classification in three groups, 40% were lower. This allowed to conclude that this solution was not the best. The contribution of the three factors obtained with the Statistical Model of Impact Measuring, in the classification of five and four groups, was significant ($P<0.001$ and $P<0.01$), according to statistical Fisher (F). Nevertheless, the classification in four groups was selected due to the wide range of factor distribution. The confirmatory analysis of K-Means allowed to improve the classification performed in the Statistical Model of Impact Measuring, because it showed higher heterogeneity among groups and homogeneity within the group, which allowed a better characterization.

Key words: *milk production, Statistical Model of Impact Measuring, method of K-means*

According to the terminology of Gondar (2003), the analysis of hierarchical cluster is known as exploratory, and the procedure of variable type representation, as profile analysis of the groups or clusters. Cuadras (2012) stated that one inconvenient of the analysis of hierarchical cluster is that it does not offer unique solutions, in spite of the existence of a real data classification structure, because the solutions depend on the considered variables and on the used method of analysis.

Hair *et al.* (1995) indicated the existence of two types of cluster analysis: hierarchical, considered as exploratory, and non hierarchical confirmatory (K-means).

Linares *et al.* (1986), Gondar (2003), Uriel and Aldás (2005) and Varela and Torres (2005) recommended the use of statistical criteria for guaranteeing the predictive validity of the formed groups, when the analysis of hierarchical classification is used, which is based on tree construction for organizing individuals in groups. Hair *et al.* (1995) proposed the use of non hierarchical techniques to confirm the results of hierarchical procedures, mainly because in the firsts, the result will depend on the amount of groups selected by the researcher at the beginning of the grouping, according to the theoretical, objective and practical knowledge.

In the Statistical Model of Impact Measuring (SMIM), Torres *et al.* (2008, 2013) classified the systems (individuals), regarding the obtained impacts. For that purpose, these authors used the hierarchical method with Euclidian distance and the grouping method of Ward, for later typifying the variables according to the formed groups. The decision of the amount of groups to be formed was taken after considering the agglomeration or dissimilarity coefficient and the judgment of the researcher.

The objective of this research was to compare the classification obtained using SMIM with the non hierarchical classification of K-Means, as a confirming method, regarding the impacts estimated in a study of 90 dairy farms from Pastaza province, Ecuador.

Materials and Methods

The study was carried out in Pastaza province, Ecuador. A total of 90 farms were analyzed as well as the corresponding variables: slope (%), farm area (ha), area used by cattle (ha), compatible area of grazing (%), number of cows (head), female cows for reproduction (head), cows for dairy (head), milk production per year (thousand L), number of gullies ha^{-1} and soil depth (cm).

The SMIM was applied and the impact indexes were used. The classifications in five, four and three groups were selected, which were contrasted with the application of Eta coefficient (Gondar 2003), which takes values from 0 to 1 and analyses the relations among the groups selected in the SMIM, after considering the variable slope, number of groups (five, four and three) and the means of original variables (10). The Eta values, close to the unit, indicate the correct group structure.

Besides, the non hierarchical cluster of K-means was used, which was identified by Hair *et al.* (1995) as a classification method for confirming and optimizing clusters. This method uses the amount of groups considered as adequate, and confirms the null hypothesis stating that the formed groups are significantly equals, regarding all and each selected factors in the SMIM. This method also determines the contribution degree of each factor to the grouping process, through the statistical test of Fisher (F).

Each procedure was processed through the statistical software IBM SPSS Statistics 22 (2013).

Results and Discussion

The association measurements Eta (table 1) showed the variables associated to the classified groups through this coefficient. For the classification in five groups, values were over 0.60, except the value of soil depth (cm) that was 0.50, and 90% of the variables were associated to the formed groups.

For the classification in five groups, the performance of Eta was similar to the selection of five groups. In this classification, the variable soil depth (cm) was inferior to 0.60, with the lowest value (0.41).

In studies carried out by Demey *et al.* (1994) and Martínez-Melo (2011), rice and milk producer farms were classified, respectively. The hierarchical cluster method was applied without other confirmatory methods of these classifications.

In the classification of three groups, six variables had values superior to 0.60: slope (%), compatible area of grazing (%), number of cows (head), cows for dairy (head), milk production per year (thousand L) and number of gullies ha^{-1}. The remaining values were very close to 0 values. It can be concluded that this is not the adequate solution.

This procedure is not subjective. The association level was higher in the classifications of five and four groups. In this case, most variables had high values of Eta coefficient, very close to one.

In the non hierarchical clustering method of K-Means, the levels of significance observed are not corrected, so, they can be considered as an evidence of the hypothesis, which states that the means of the groups are equal. This confirmatory analysis was performed equally with the impact indexes of the three factors selected in the SMIM.

Table 2 and 3 show the results of the ANAVA, obtained by the classification method of K-Means, for organizing them into five and four groups, where the impact indexes of the three factors estimated through the SMIM took a significant part (P<0.001) of the grouping process.

The highest contributing factors to the grouping process, in order of importance, were: environmental situation, usage size, herd and production, variables that are included in these factors.

Once confirmed the significance of impacts on group formation, there was no determination of contribution degree of each one to the grouping process. If the value resulting from the statistical test F increases, their contribution increases too. The results showed that the factors environmental situation, usage size, herd and production had the same order in their contribution to the classification in five and four groups.

These contributions for the three factors were: 51.56, 31.45 and 28.14 in five groups, and 73.81, 33.50 and 13.56 in four groups. This last one showed a higher range of contribution, which confirmed that it is the

Table 1. Eta coefficient for classifying farms in five, four and three groups

Variables	Eta		
	Five groups	Four groups	Three groups
Slope , %	0.79	0.79	0.78
Farm area, (ha)	0.78	0.72	0.51
Area used by cattle, (ha)	0.79	0.75	0.54
Compatible area of grazing, %	0.78	0.78	0.77
Number of cows, heads	0.63	0.63	0.62
Number of female cows for reproduction, heads	0.60	0.60	0.59
number of cows for dairy, head	0.68	0.67	0.66
Milk production per year, thousand L	0.63	0.61	0.61
Gullies ha^{-1}	0.77	0.75	0.75
Soil depth, cm	0.50	0.41	0.41

Tabla 2. Contribution of variables to the formation of the five groups

Factors	Cluster		Error		F	Sig.
	Square Mean	Degree of freedom	Square Mean	Degree of freedom		
Herd and production impact indexes	12.677	4	0.451	85	28.139	0
Environmental situation impact indexes	15.756	4	0.306	85	51.563	0
Usage size impact indexes	13.279	4	0.422	85	31.456	0

Tabla 3. Contribution of variables to the formation of the four groups

Factors	Cluster		Error		F	Sig.
	Square Mean	Degree of freedom	Square Mean	Degree of freedom		
Herd and production impact indexes	9.528	3	0.703	86	13.563	0
Environmental situation impact indexes	21.368	3	0.289	86	73.811	0
Usage size impact indexes	15.987	3	0.477	86	33.503	0

best because it formed higher heterogeneity among groups. Benítez (2007) classified and typified five macizos montañosos in Cuba through the application of combined multivariate methods, which include the use of hierarchical cluster.

Vargas *et al.* (2013), using the analysis of main components and weight factors, characterized and classified milk farms. However, they did not use the Eta coefficient to confirm the formed groups, but to contrast them using the univaried analysis of variance for each variable. These authors used the non hierarchical cluster method of K-Means to optimize the number of groups.

Table 4 shows the type representation of variables, according to the groups formed in the classification considered as the best with K-Means. The obtained distribution was different from the one obtained with SMIM. Values between parentheses indicate the number of farms in each group of SMIM.

The first group was composed by 51 farms: 47 belong to the group I of SMIM, and four to the group II. The second group, with 31 farms, was composed by 22 farms from group II, two from the group III and seven from group IV. The third group had six farms: five from group III and one from group IV, all from SMIM. The forth group was composed by two farms from group III.

The first group was characterized by farms with low slope, small size and percentage of area compatible with the highest grazing. The number of cows, reproducer and milking cows, agreed with the size of the farms and the area destined to grazing. These farms had lower amount of gullies ha^{-1} and soil depth was superior to the remaining groups.

The second group was formed by farms located in a superior slope and the area compatible with grazing was much lower. The amount of gullies ha^{-1} was three times higher than those of group I.

Farms from third group had larger size than the previous groups. They had superior number of cows, milking cows and reproducers. Therefore, milk production (thousand L) was superior. The number of gullies ha^{-1} was between group I and group II.

The farms from group four had the highest slope, the largest area and the largest amount of area used by cattle. However, they had lower surface compatible with grazing. Also, these farms had the lowest number of reproducers and milking cows, with the lowest milk production, regarding the remaining farms from the other groups. These farms also had the highest number of gullies ha^{-1}.

The confirmatory analysis (K-Means) allowed to improve the classification of farms, according to SMIM, because the groups were distributed more heterogeneously, which allowed to obtain a better characterization of these groups.

Tabla 4. Type representation of groups formed by the K-means method in the formation of four groups

Variables	51 farms (47)		31 farms (26)		6 farms (8)		2 farms (9)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Slope , %	17.8	10.3	49.8	13.4	30.0	9.5	52.5	10.6
Farm area, (ha)	44.6	23.1	43.0	21.5	134.5	60.1	175.0	35.4
Area used by cattle, (ha)	33.6	17.6	31.1	14.5	100.7	27.7	130	28.3
Compatible area of grazing, %	84.6	14.8	36.1	23.2	68.3	9.8	35.0	7.1
Number of cows, heads	15.1	6.8	15.6	8.6	36.5	15.5	21.0	12.7
Number of female cows for reproduction, heads	17.2	8.0	19.0	11.2	40.2	19.5	9.0	7.1
number of cows for dairy, head	8.9	3.3	10.5	6.0	22.3	3.9	6.5	0.7
Milk production per year, thousand L	21.6	13.1	26.3	18.6	61.9	30.8	17.3	11.7
Gullies ha^{-1}	35.1	26.0	107.4	33.6	63.2	17.8	105.0	7.1
Soil depth, cm	21.4	8.5	15.4	3.9	18.7	5.7	14.5	7.8

References

Benítez, D. G. 2007. El manejo de la finca ganadera en la montaña. Bayamo-Cuba. Instituto de Investigaciones Agropecuarias "Jorge Dimitrov". Granma, Cuba

Cuadras, C. M. 2012. Nuevos Métodos de Análisis Multivariado. Barcelona-España. CMC Editorial Monacor.

Demey, J. R., Adams, M. & Freites, H. 1994. Uso del método de análisis de componentes principales para la caracterización de fincas agropecuarias. Agronomía Tropical. 44:475

Gondar, J.E. 2003. Metodología de la investigación estadística con SPSS. Madrid. Ed. Data Mining Institute S.L.

Hair, J., Anderson, R. & Tatham, R. 1995. Multivariate Data Analysis. 4° Ed. Englewood Cliffs: Prentice Hall

IBM SPSS. 2013. IBM SPSS Statistics 22. Algorithms. Chicago: IBM SPSS Inc.

Linares, G., Acosta, L. & Sistach, V. 1986. Estadística Multivariada. La Habana. Universidad de La Habana, Cuba. 319 pp.

Martínez-Melo, J., Jordán, H., Torres, V., Fontes, D., Lezcano, Y.& Cubillas, N., 2011. Classification of dairy units belonging to the Basic Units of Cooperative Production in Ciego de Avila, Cuba. Cuban J. Agric. Sci. 45:373

Torres, V., Cobo, R., Sánchez, L.& Raez, N. 2013. Statistical tool for measuring the impact of milk production on the local development of province in Cuba. Livestock Res. Rural Development 29 (9)

Torres, V., Ramos, N., Lizazo, D., Monteagudo, F. & Noda, A. 2008. Statistical model for measuring the impact of innovation or technology transfer in agriculture. Cuban J. Agric. Sci. 42:13

Uriel, E. &Aldás, J. 2005. Análisis Multivariante Aplicado. Madrid España. Thomson Editores Spain.

Vargas-Leitón, B., Solís-Guzmán, O., Sáenz-Segura, F. & León-Hidalgo, H. 2013. Caracterización y clasificación de hatos lecheros en Costa Rica mediante análisis multivariado. Agronomía Mesoamericana 24:257

Varela, M. & Torres, V. 2005. Application of three-mode principal components analysis in the multivariate characterization of king grass somaclones. Cuban J. Agric. Sci. 39:527

Received: June 15, 2014