PLURIS 2008

809

MULTI-CRITERIA SPATIAL ANALYSIS WITH MACHINE LEARNING ALGORITHM - AN APPLICATION IN THE SOUTH OF BRAZIL

Rui António Rodrigues Ramos
rui.ramos@civil.uminho.pt

Rochele Amorim Ribeiro
rochele@sc.usp.br

ENDEREÇO PARA CORRESPONDÊNCIA:
Rochele Amorim Ribeiro
Universidade de São Paulo - Escola de Engenharia de São Carlos
Departamento de Transportes
Av. Trabalhador São-carlense, 400
13.566-590 Centro São Carlos - SP - Brasil

RESUMO

This paper explores a multicriteria spatial analysis methodology with a machine learning algorithm, the Classification Tree Analysis (CTA) within Idrisi GIS, to classify and identify homogeneous regions. The proposed approach is tested in a case study carried out in the South of Brazil. All the municipalities were classified and grouped within areas according to similar condition of urban preponderance in socioeconomic and environmental indicators. The results are evaluated and compared with two other methodologies previously implement by the authors: (a) a ranking of municipality through an aggregate index; and (b) using Kohonen´s Self-Organizing Map (SOM) as an unsupervised classifier. The identification of similar areas with analogous socioeconomic and environmental characteristics is important to the development of regional and municipal common sustainable strategies and advances in municipality partnerships.

MULTI-CRITERIA SPATIAL ANALYSIS WITH MACHINE LEARNING ALGORITHM – AN APPLICATION IN THE SOUTH OF BRAZIL

R.A.R. Ramos, R. A. Ribeiro

ABSTRACT

This paper explores a multicriteria spatial analysis methodology with a machine learning algorithm, the Classification Tree Analysis (CTA) within Idrisi GIS, to classify and identify homogeneous regions. The proposed approach is tested in a case study carried out in the South of Brazil. All the municipalities were classified and grouped within areas according to similar condition of urban preponderance in socioeconomic and environmental indicators. The results are evaluated and compared with two other methodologies previously implement by the authors: (a) a ranking of municipality through an aggregate index; and (b) using Kohonen´s Self-Organizing Map (SOM) as an unsupervised classifier. The identification of similar areas with analogous socioeconomic and environmental characteristics is important to the development of regional and municipal common sustainable strategies and advances in municipality partnerships.

1 INTRODUCTION

Presently, the sustainable development challenges take into account the social and economic background. Therefore, decision makers in central government, regional development agencies and local communities promote similar sustainable goals to regular territories while protecting the environment. However, in practice, this process is difficult to be implemented because the identification and aggregation of municipalities with similar characteristics is not a simple political act and commonly is supported by subjective criteria. For this reason, in the last decades, urban and regional studies tried to enlarge the application of emergent planning techniques, for example, the use of geoprocessing toolbox, spatial statistics and neural networks in a geographic information system. Hence, through these techniques of processing and analyzing spatial data in several information layers, it is possible to consider multidisciplinary characteristic of decisions in regional and urban planning. So, it is possible to evaluate territorial scenarios with socioeconomic and environmental features and spatial neighborhood relationships. The scenarios can be developed by combining visualization and data analysis in a geographical information system (GIS).

The aim of the paper is to explore a multicriteria spatial analysis methodology with a machine learning algorithm, the Classification Tree Analysis (CTA) within Idrisi GIS, to classify and identify territories through demographics, socioeconomic and environmental indicators.

The proposed approach is tested in a case study carried out in the South of Brazil, and the study area is shaped by the states of Paraná, Santa Catarina e Rio Grande do Sul. The

territorial extension of the study area is 500 000 km^2, occupying 6.8% of Brazilian territory. The total population is 25 millions of inhabitants, approximately 15% of Brazilian population. All the 1159 municipalities were classified and grouped according to similar condition of urban preponderance in demographic, socioeconomic and environmental indicators.

The results obtain with the implementation of the Classification Tree Analysis are evaluated and compared with two methodologies applied by the authors in a previous work (Ribeiro & Ramos, 2007): (a) a ranking of municipality through an aggregate index; (b) Kohonen´s Self-Organizing Map (SOM) neural network as an unsupervised classifier.

The identification of similar areas, with analogous socioeconomic and environmental characteristics, constitutes an option to avoid the confrontations derived from the mainly subjective political criteria, turning important to the development of common sustainable strategies and the influence in advances in municipality partnerships.

2 CLASSIFICATION TREE ANALYSIS

Techniques for combining multiple information sources are growing in remote sensing research areas, mainly in landscape mapping (Rogan & Miller, 2006). Hence, the machine-learning techniques are rising because they facilitate the integration of data from several sources due to their ability to combine continuous and categorical data analyses in statistical assumptions (Gahegan, 2003).

Machine-learning refers to induction algorithms that analyze information and recognize patterns through automated and repeated learning processes from training data (Malerba et al., 2001). Breiman et al. (1984) presents an example of machine-learning techniques as a classification tree analysis, which is capable to deal with a high dimensional data. This technique has been widely applied in vegetation modeling from environmental GIS data (Miller & Franklin, 2002) and in image analysis of land-cover and forest mapping (Friedl & Brodley, 1997; Rognan et al., 2003).

At the moment, it is possible to find classification decision tree algorithm integrated within GIS software, like the CTA module in the IDRISI GIS software. In this case, CTA is a machine learning algorithm who classifies and analyzes raster databases. This is an analytical procedure that takes examples of known classes (i.e., training sites) and constructs a decision tree based on measured attributes. Because it takes a set of training data and constructs a decision tree, CTA is a form of machine learning, like a neural network. However, unlike a neural network, CTA produces a white box rather than a black box solution because the nature of the learning process is a discrete output (Eastman, 2006).

In the CTA module of IDRISI GIS software, the CTA process must start using the training data to build a classification tree. Through the training site information, the binary splitting rule is identified as a threshold in one of the multiple input images that isolates the largest homogenous subset of training pixels from the remainder of the training data. Basically, the algorithm selects the attribute (associated to pixel data) and value that divides a set of samples into two groups, minimizing the variability within each subgroup while maximizing the contrast between the groups. The tree grows by splitting data at each internode into new internodes containing progressively more homogeneous sets of training

pixels. A newly grown internode may become a leaf when it contains training pixels from only one class, or pixels from one class dominate the population of pixels in that internode, and the dominance is at an acceptable level specified by the user (significant level). When there are no more internodes to split, the final classification tree rules are formed.

As refers Thuiller et al. (2003), Regression and Classification Trees provide an alternative to regression techniques. They do not rely on a priori hypotheses about the relation between independent and dependent variables. This method consists of recursive partitions of the dimensional space defined by the predictors into groups that are as homogeneous as possible in terms of response. The tree is built by repeatedly splitting the data, defined by a simple rule based on a single explanatory variable. At each split, the data are partitioned into two exclusive groups, each of which is as homogeneous as possible. The method builds a nested sequence of subtrees by recursively snipping off the less important splits in terms of explained deviance. The length of the tree was controlled by choosing the best trade-off between explained deviance and tree size. Each predictor could be used several times in the tree if it improved the predictive performance.

In the present work we assessed not remote sensing data but demographic, socioeconomic and environmental indicators data from municipalities in terms of their characteristics in sustainability dimensions. So, the paper used the Classification Tree Analysis to test whether the municipalities could be correctly assigned to four groups from which they were sampled. This analysis is a nonparametric technique that uses a recursive-partitioning algorithm to repeatedly partition the dataset into a nested series of mutually exclusive groups, each of them as homogeneous as possible with respect to the response variable (here, indicators for each sustainable dimension).

3 METHODOLOGY

The methodology structure of the work was elaborated in three parts:

i. Applying the CTA supervised classification for identifying municipality groups with homogeneous characteristics. This process was done in two different manners:

   a. the Ranking and SOM classification groups for all municipalities are used as training sites. These processes we are going to call CTA_Rank and CTA_SOM respectively;

   b. the Ranking and SOM classification groups are used as training site but only with a sample of municipalities. These processes we are going to call CTA_sel_Rank and CTA_sel_SOM respectively.

ii. Comparing the results to find significant differences between groups classification;

iii. Evaluating the results by means of a qualitative analysis.

The data base and the techniques of data treatment will be explained in following section.

4 CASE STUDY

4.1 Data base characterization

In a previous work, Ribeiro & Ramos (2007) used two data bases: one geographic data base with boundaries of Brazil's southern municipalities, and another data base with demographic, socioeconomic and environmental municipality indicators (see Table 1). The geographic data base was obtained from the Territorial Unit Mapping of the Brazilian Institute of Geographic and Statistic (IBGE, 2001). The data base indicators for the municipalities were extracted from Human Development Atlas of Brazil (PNUD, 2003).

Table 1 List of Indicators for each sustainable dimension

Dimensions	Indicators
Demographic	D1 - Literacy Rate (%)
	D2 - Municipal Human Health Development Index (*)
	D3 - Municipal Human Education Development Index (*)
Socioeconomic	SE1 - Percentage of population who lives in homes with private car
	SE2 - Percentage of population who lives in homes with public electric energy supply
	SE2 - Municipal Human Income Development Index (*)
Environmental	E1 - Percentage of population who lives in homes with public water supply
	E2 - Percentage of population who lives in homes with waste collection service

(*) The index values were multiplied by 100, for suiting with the percentages. Thus, the scale of 0-1 becomes a scale of 0-100.

Therefore, all the municipalities were classified in four groups with similar characteristics. The order of the groups represents a sustainability level according to values of municipality indicators. Due to this, the Group one represents the municipalities with the best score for sustainability; and, on the other hand, the Group four represents the municipalities with worse scores. In this previous work the authors used two methods to classify the municipalities:

(i) a ranking method, which applies a municipality classification through an aggregator index. The aggregator index integrates the information of all indicators. The municipalities were classified in four groups with similar characteristics based on the aggregator index ranking. The ranking is organized in decreasing order of values and the group delimitation is based on average and standard deviation values obtained from the aggregator index;

(ii) the Kohonen´s Self-Organizing Map (SOM), within IDRISI GIS software, which processes an arrangement of the municipalities through a neural networks with an unsupervised classification techniques. As input data were used the indicators values associated with each municipality, resulting in an output of four classification groups.

In the present work, the data base is the same used before (Ribeiro & Ramos, 2007), organized in indicator data and group classification for both methods, as follow:

a) Geographic raster images with information of data base indicators (listed in Table 1) for each municipality;

b) Geographic raster images with Ranking and SOM classification groups (represented in Figure 1 and 2).

The methodology that processes these data base will be discussed in the next subsections.

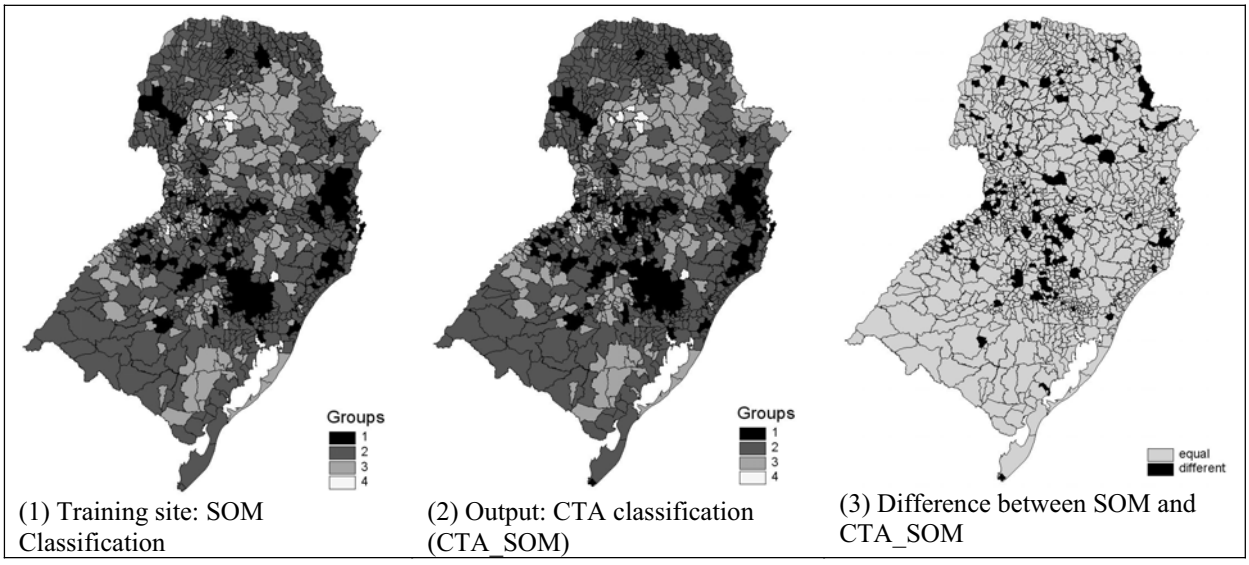
4.2 Applying the CTA supervised classification

Because CTA is a supervised classification it is necessary to prepare data with the training site to define the rules of the decision tree. So, for the training site, the Ranking and SOM classification outputs was assumed. In the same way, for the input data the geo-information indicators of the municipalities were used. The IDRISI GIS software (Idrisi 15.0 – Andes Edition, Clark Labs) was the platform for the analysis.

Firstly, CTA classification was applied using as training site the information of all municipalities. The classification results are shown in

Fig. 1,

Fig. 2 and Table 2.

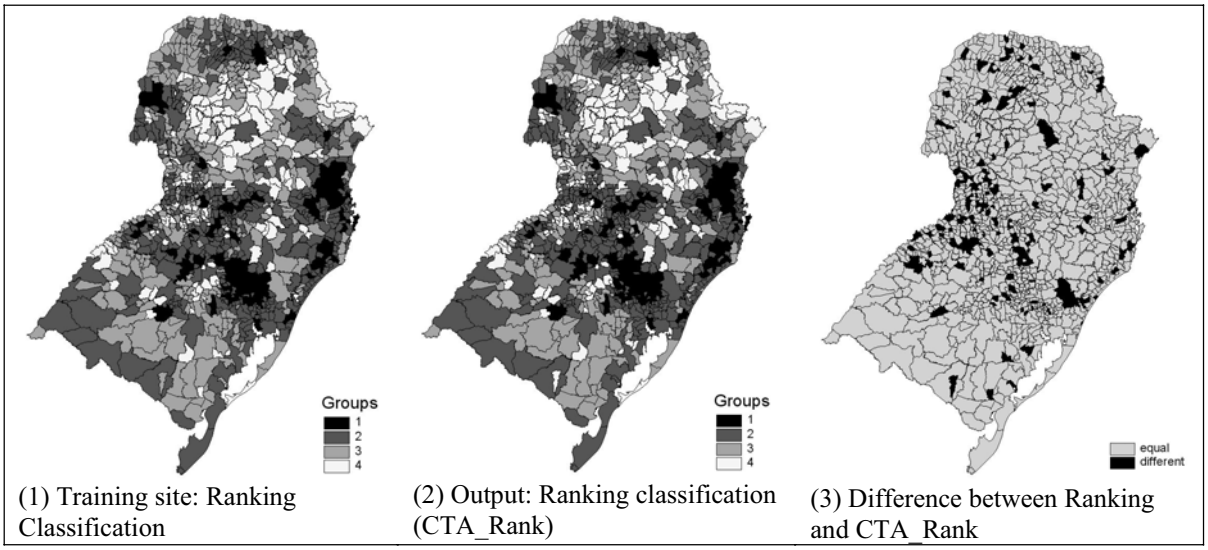


(1) Training site: SOM Classification

(2) Output: CTA classification (CTA_SOM)

(3) Difference between SOM and CTA_SOM

Fig. 1 CTA classification with SOM output as training site



(1) Training site: Ranking Classification

(2) Output: Ranking classification (CTA_Rank)

(3) Difference between Ranking and CTA_Rank

Fig. 2 CTA classification with Ranking output as training site

Table 2 Differences between CTA classification and Ranking/SOM classifications

	Number of municipalities *method*		
group	rank	cta_rank	difference
1	172	181	9
2	453	464	11
3	365	322	-43
4	169	192	23

	Number of municipalities *method*		
group	som	cta_som	difference
1	206	241	35
2	657	623	-34
3	287	283	-4
4	9	12	3

Afterwards, a sample of municipality was used as training site for processing CTA classification. In each classified group, a sample of 25% of municipalities was selected randomly three times, resulting in three different samples. The CTA classification was tested using these samples and the difference among them was evaluated. The difference between the results was not significant (1% level). Hence, only one of these results will be compared with other methods.

The amount of municipalities selected in each method is shown in Table 3. The training site result was shown in

Fig. 3. The output CTA process and difference between methods are shown in

Fig. 4 and

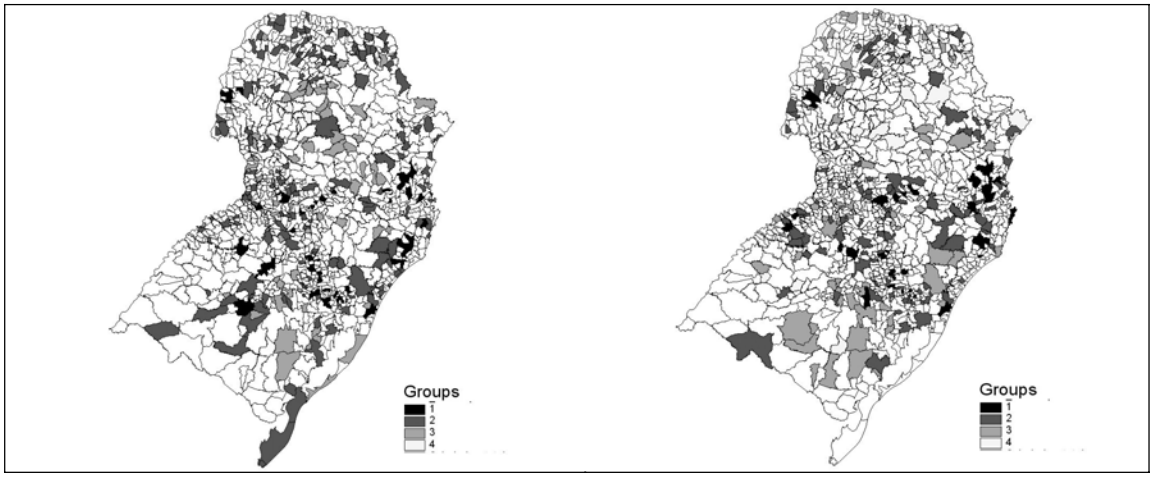
Fig. 5. The comparison of the number of municipalities between CTA_sel_Rank / CTA_sel_SOM and Ranking/SOM methods is shown in Table 4.

Table 3 Amount of municipalities selected in each method for sampling

RANK	**sample seletion (25%)**	
group	**total**	**nº mun.**
1	172	43
2	453	114
3	365	92
4	169	43

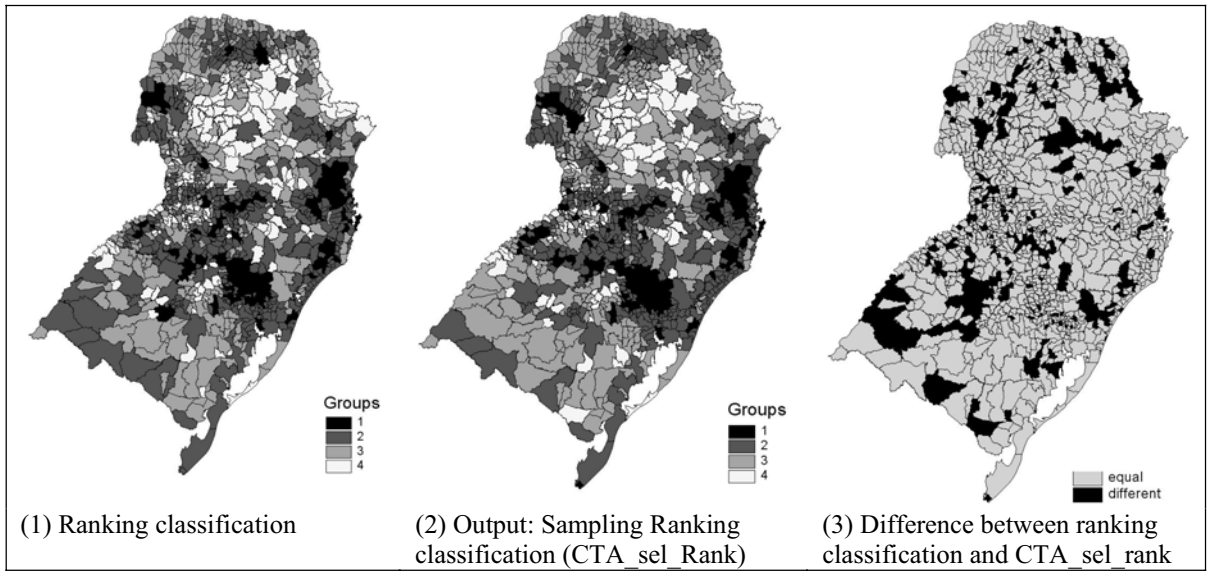
SOM	**sample seletion (25%)**	
group	**total**	**nº mun.**
1	206	52
2	657	165
3	287	74
4	9	9*

*all was selected



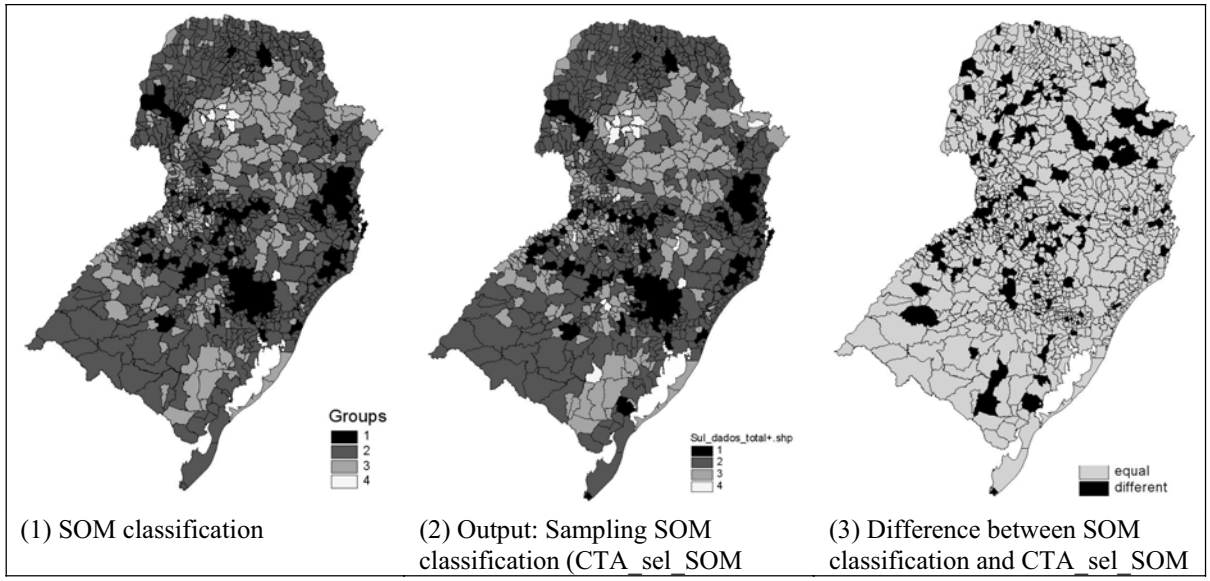
(1) Training site – sample of municipality of SOM classification	(2) Training site – sample of municipality of Ranking classification

Fig. 3 Results of sample of municipality selection



(1) Ranking classification | (2) Output: Sampling Ranking classification (CTA_sel_Rank) | (3) Difference between ranking classification and CTA_sel_rank

Fig. 4 CTA classification with training site municipalities' sample selection of Ranking



(1) SOM classification | (2) Output: Sampling SOM classification (CTA_sel_SOM | (3) Difference between SOM classification and CTA_sel_SOM

Fig. 5 CTA classification with training site municipalities' sample selection of SOM

Table 4 Difference between CTA_sel_Rank/CTA_sel_SOM classifications and Ranking/SOM classifications

	Number of municipalities				Number of municipalities		
	method				*method*		
grupo	rank	cta_sel_rank	difference	**group**	SOM	cta_sel_SOM	difference
1	172	187	15	*1*	206	198	-8
2	453	425	-28	*2*	657	647	-10
3	365	391	26	*3*	287	296	9
4	169	156	-13	*4*	9	18	9

4.3 Results Analysis

For comparing the classification results, the Analysis of Variance (ANOVA) statistic test was applied. ANOVA is a method used to test differences among sample means. It is a general test, which permits testing two samples or more. ANOVA compares the variation due to the experiment treatments and the random variation. The goal is to determine if the difference among means depends on the treatment or is random. So, this study proposes to verify if there is a significant difference between the methods (here methods is equal to treatments). The test compared the difference among groups, among indicators and among methods. Five comparisons were made: i) comparison between SOM classification and CTA_SOM; ii) comparison between Ranking classification and CTA_Rank; iii) comparison between SOM classification and CTA_sel_SOM; iv) comparison between Ranking classification and CTA_sel_Rank; v) comparison between CTA_SOM and CTA_rank.

The results showed that there are not significant differences among four of methods in comparisons (i),(ii), (iii) and (iv); however, the difference in comparison (v) is significant. The differences among groups and among indicators are significant in all comparisons.

It is important to identify differences among groups and among indicators. The difference among groups indicates that the method proposed could classify efficiently four distinct homogeneous regions. The difference among indicators shows that each item of the indicator data base represents discrete influences in classification processes. The indicators characterize the municipalities in specific ways representing a wide sustainable spectrum. On the other hand, differences among methods are not desirable in experiments (i), (ii),(iii),(iv) because the CTA method should be similar to the method that was used as training site classification (rank or SOM methods). But, in the experiment (v) a difference among methods was expected because this result agrees with Ribeiro & Ramos (2007), whose conclusions discuss the dissimilarity between Ranking and SOM methods.

The statistical parameters and results of the ANOVA test are shown in Table 5. Graphics showing the factors means are present in

Fig. 6. In these graphics, similarities and discrepancies among groups and indicators in each method are easy to identify.

Table 5 Results of ANOVA test in each comparison

Comparison	*Squared multiple R*	*Factors*	*F-ratio*	*p-value*	*Significant difference*
(i) Rank x CTA_rank	0.973	Group	72.055**	0.000	Yes – at 1% level
		Indicators	225.938**	0.000	Yes – at 1% level
		Method	0.020	**0.887**	**No**
(ii) SOM x CTA_SOM	0.969	Group	19.942**	0.000	Yes – at 1% level

		Indicators	221.275**	0.000	Yes – at 1% level
		Method	0.003	**0.956**	**No**
(iii) Rank x CTA_sel_rank	0.972	Group	72.055**	0.000	Yes – at 1% level
		Indicators	225.938**	0.000	Yes – at 1% level
		Method	0.020	**0.887**	**No**
(iv) SOM x CTA_sel_som	0.928	Group	59.189	0.000	Yes – at 1% level
		Indicators	70.613	0.000	Yes – at 1% level
		Method	0.482	**0.490**	**No**
(v) CTA_rank x CTA_SOM	0.929	Group	41.018**	0.000	Yes – at 1% level
		Indicators	77.928**	0.000	Yes – at 1% level
		Method	9.355*	**0.004**	**Yes – at 5% level**

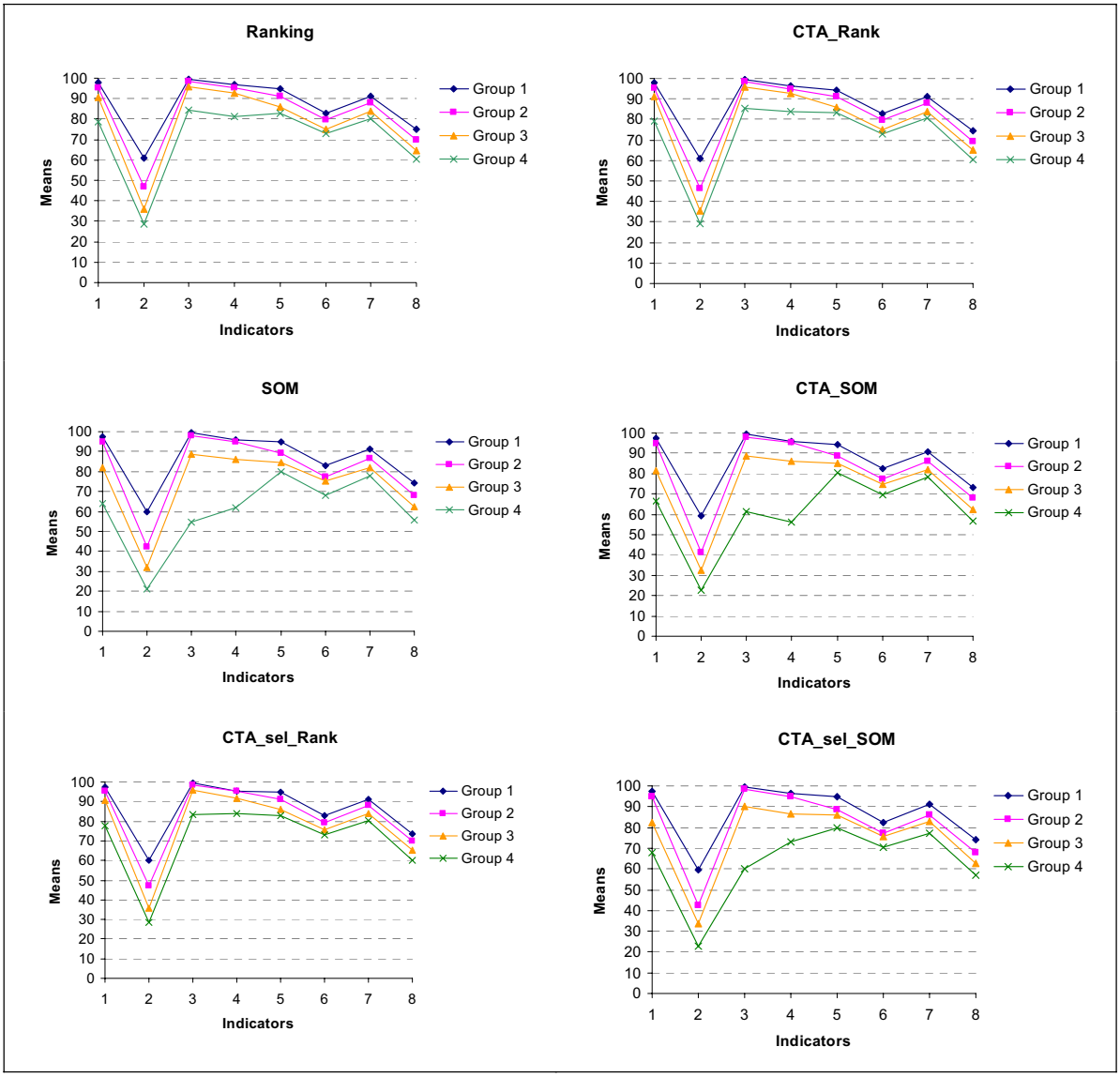


Fig. 6 Indicator behavior in each method for each group.

5 CONCLUSIONS

The results obtained through the CTA emphasize that SOM and Ranking methods are significantly consistent. Therefore, these results can be considered a step forward from the previous work (Ribeiro & Ramos, 2007). Using full data or only 25% random sample data the CTA method produces the same classification output of the Ranking or SOM methods

used as training site. This fact reinforces that the models proposed in the previous work for classifying homogenous regions are valid and balanced in terms of identifying municipality classification with in sustainable dimensions indicators.

This is a promising approach for analyzing and integrating GIS data concerning the interpretation of several indicators in a geospatial model, not only in remote sensing data but also in geographical data base for political delimitation proposals.

6 BIBLIOGRAPHY

Breiman, L.; Friedman, J.H.; Olshen, A.; Stone, C.G. (1984) **Classification and Regression Trees**. Wadsworth International Group: Belmont, California, USA.

Eastman, J.R. (2006) **IDRISI Andes: Guide to GIS and Image Processing**. Clark Labs, Clark University, Worcester, Massachusetts, USA.

Friedl, M.A. and Brodley, C.E. (1997) Decision tree classification of land cover from remotely sensed data. **Remote Sensing of Environment** 61(3): 399-409.

Gahegan, M. (2003) Is inductive machine learning just another wild goose (or might it lay the golden egg)?. **International Journal of Geographical Information Science** 17: 69–92.

IBGE – Instituto Brasileiro de Geografia e Estatísca. Geociências (2001) **Malha Municipal Digital de 2001**. Available online at: <http://www.ibge.gov.br/home/geociencias/default_prod.shtm#TERRIT>.

Malerba, D.; Esposito, F.; Lanza, A.; Lisi, F.A. (2001) Machine learning for information extraction from topographic maps. In H. J. Miller & J. Han (Eds.), **Geographic Data Mining and Knowledge Discovery**. New York, NY: Taylor and Francis, Inc.

Miller, J. and Franklin, J. (2002) Modeling the distribution of four vegetation alliances using generalized linear models and classification trees with spatial dependence. **Ecological Modelling** 157: 227-247.

PNUD - Programa das Nações Unidas para o Desenvolvimento (2003) **Atlas do Desenvolvimento Humano no Brasil**. Instituto de Pesquisas Econômicas Aplicadas, Fundação João Pinheiro. Available online at: <http://www.pnud.org.br/atlas/dl/único /AtlasIDH2000.exe > .

Ribeiro, R.A. and Ramos, R.A.R. (2007) A methodological approach in order to define a geographical territorial structure classification through an application in the South of Brazil. In: **10th International Conference on Computers in Urban Planning and Urban Management - CUPUM 2007**, Brazil.

Rogan, J., and Miller, J. (2006) Integrating GIS and Remotely Sensed Data for Mapping Forest Disturbance and Change. In M.A. Wulder and S.E. Franklin (Eds.), **Understanding Forest Disturbance and Spatial Pattern: Remote Sensing and GIS Approaches**, Taylor and Francis.

Rogan, J.; Miller, J.; Stow, D.; Franklin, J.; Levin, L.; Fischer, C. (2003) Land-cover Change Monitoring with Classification Trees Using Landsat TM and Ancillary data. **Photogrammetric Engineering & Remote Sensing** 69(7): 793-804.

Thuiller, W.; Araújo, M.B.; Lavorel, S. (2003) Generalized models vs. classification tree analysis: predicting spatial distributions of plant species at different scales. **Journal of Vegetation Science** 14: 669-680