Development of dissimilarity functions using stochastic distances for region-based land cover classification: a case study near Tapajós Flona, Pará state, Brazil.

Luciano Vieira Dutra [1]
Rogerio Galante Negri [2]
Sidnei João Siqueira Sant´Anna[1]
Dengsheng Lu [3]

[1]Instituto Nacional de Pesquisas Espaciais - INPE
Caixa Postal 515 - 12245-970 - São José dos Campos - SP, Brasil
{dutra, sidnei}@dpi.inpe.br

[1] Universidade Estadual Paulista – UNESP
Instituto de Ciência e Tecnologia - ICT
Rodovia Pres. Dutra, km 137,8 - 12247-004 - São José dos Campos - SP, Brasil
rogerio.negri@ict.unesp.br

[3] Michigan State University – MSU
Center for Global Change and Earth Observations
East Lansing – Michigan, USA
ludengsh@msu.edu

Abstract. One recent alternative to standard pixel based classification of remote sensing data, is the region based classification (RBC), which has been proved particularly useful when analyzing high resolution imagery of complex environments, like urban areas. First the imagery is decomposed into homogenous regions, following some criteria, and then each region is classified to one of the classes of interest. Normally, classification is performed by using stochastic distances, which measures the distance of the pixels distribution inside an unknown region and the representative distributions of each class. The class, whose distance is minimum to the unknown region distribution, is assigned to the region, which is known as stochastic minimum distance classification (SMDC). A problem appears when one, or more, class distribution is multi-modal, which violates the Gaussian hypotheses used for classes distributions, degrading the mapping accuracy. This investigation reports the usage of different compositions of the original stochastic minimum distance classifier with the objective of getting less sensitive results for classification, when potentially multi-modal classes are used. The newly developed classifier, called stochastic nearest distance classifier (SNDC), produced the best result when compared with the original classifier and other possible compositions, in a study case near the Tapajós Flona, in Pará state, Brazil. This study also brings, as methodological contribution, a criterion to improve the segmentation phase of RBC methods.

Keywords: Remote Sensing, Image Processing, Geology.

1. Introduction

Region-based classification (RBC) methods has been increasingly used, particularly when using high resolution imagery acquired over urban areas, where per point classification normally fails, because of high heterogeneity and complexity of such environments. (Liu and Xia 2010). RBC is also especially useful for using with radar data, which is normally analyzed with per point methods.(Li et al., 2012; Lu et al.,2011). Region-based classifiers first aggregate pixels into homogeneous objects using segmentation techniques and then classify the objects individually. Normally, classification is performed by using a statistical distance of the representative distributions of each class of interest and the pixels distribution inside an unknown region. The class, whose distance is minimum to the unknown region distribution, is assigned to the region, which is known as stochastic minimum distance classification (SMDC). Gaussian assumption (Richards e Jia, 2005) is used for the standard statistical

distance definition, which implies the unimodality assumption. Many cases, in a supervised classification task, classes with multimodal distribution are collected as reference and this violation of this assumption degrades the resulting mapping accuracy.

Negri et al. (2012) theoretically introduced distinctive ways of using stochastic distances, for region-based classification, which have been tested with image data simulation. These methods are supposed to be less sensitive to multimodality. In this investigation, a practical evaluation of the proposed method is presented for land cover classification using LANDSAT-5 TM imagery in a study area near the Tapajós National Forest western part of Pará state, Brazil. Several multimodal classes, considering optical data, have been chosen to test the proposed method. Another contribution is a proposal of a systematic way of chosen which channels, from a given data set, should be used as input to the segmentor of choice.

2. Methodology and Materials.

Figure 1 presents the processing chain:

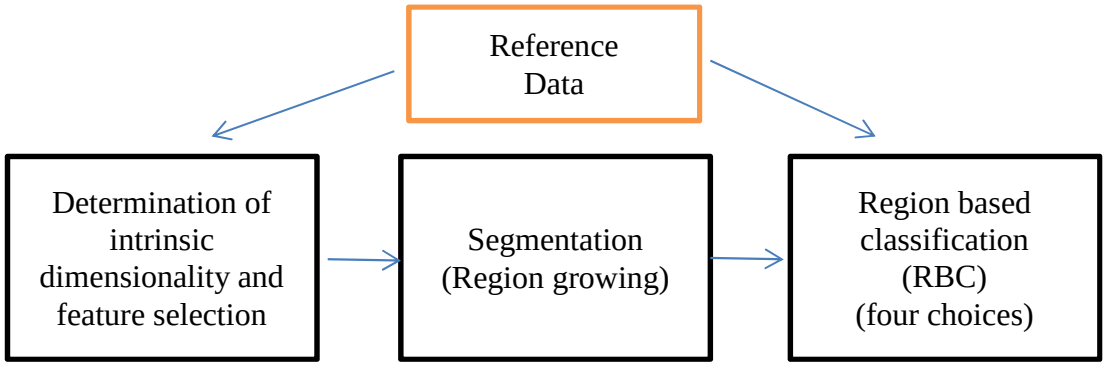

Figure 1: Processing chain for Region Based Classification

Initially reference data is collected for all classes of interest, then all data for each class will inform a feature selection technique to choose which sub-set of input channels will be used for segmentation. After segmentation, each region, as a whole entity, will be classified to one of defined classes(Section 2.3). Reference data will be used for estimating the necessary parameters for all classifiers.

Section 2.1 presents one standard stochastic distance, known as JM distance, its formulation using Bayesian assumption, and how it is normally used as a region based classifier. Section 2.2 will present three different dissimilarity classifiers constructed with alternative compositions of the JM distance.

2.1 Classifying using the Jeffries-Matusita distance.

If I is an image defined on a support $S \subset N^2$ and X is a feature space, $I(s) = \mathbf{x}$ denotes that a pixel $s \in S$ of I has an attribute vector $\mathbf{x} \in X \subset R^n$. The region-based classification process consists of associating a class $\omega_j \in \Omega,\ j = 1,\ldots,c$, to a region $R_i \subset S$, $i = 1,\ldots,r$. Ω is a set of classes of interest, R_i is a set of connected pixels s_a, $a = 1,\ldots,\#R_i$, where the attributes of s_a are obtained from $I(s_a)$ and $\#$ is the cardinality of the operator. In this context, the support of I is partitioned into r disjoint regions by a segmentation process.

Regions represent sets of spatially connected pixels whose attribute vectors meet a particular uniformity criterion. In the classification process, all pixels of the same region are associated to one class.

For a supervised region-based method, it is necessary to acquire a set of labeled regions $\mathsf{D} = \{(R_i, \omega_j) \in \mathsf{S} \times \Omega : i = 1, \ldots, m;\ j = 1, \ldots, c\}$, where m is the number of training regions. The notation (R_i, ω_j) indicates that R_i is assigned to the class ω_j. Class distributions are modeled upon information drawn from D.

Classification of an unknown region is performed associating this region to the least dissimilar class. Dissimilarity functions are any class of functions of two objects that returns zero when two objects are equal and go increasingly positive as the objects are more and more different (Theodoridis e Koutroumbas, 2008). Formally, if we let R_i be an unlabeled region and $\mathbf{M}(f_{R_i}, f_{\omega_j})$ be a dissimilarity function between the distributions f of the attribute vectors of the pixels in R_i and the class ω_j; an assignment (R_i, ω_j), is made when the following rule (Equation 1)is satisfied:

$$(R_i, \omega_j) \Leftrightarrow j = \arg\min_{j=1,\ldots,c} \mathbf{M}(f_{R_i}, f_{\omega_j}). \quad (1)$$

Stochastic distances have been used as a dissimilarity measure. These distances quantify the separability between two probability density functions that are used to model the information for a certain class or distribution of pixels values inside a region. The Jeffries-Matusita distance (**JM**) is defined by Equation 2.

$$\mathbf{JM}(C, D) = \int \left[\sqrt{f_C(\mathbf{x}; \Theta_C)} - \sqrt{f_D(\mathbf{x}; \Theta_D)}\right]^2 d\mathbf{x}, \quad (2)$$

where f_C and f_D are probability density functions, with parameters Θ_C and Θ_D, which model the information distribution of the sets C and D; such elements belong to X.

Assuming f_C and f_D are Gaussian Multivariate distributions, Equation 2 can be reformulated to (Equation 3):

$$\mathbf{JM}(C, D) = 2\left(1 - e^{-\mathbf{B}_G(C,D)}\right), \quad (3)$$

where $\mathbf{B}_G(\cdot,\cdot)$ is the Bhattacharyya distance under assumption of the Gaussian Multivariate distribution, and is defined by:

$$\begin{aligned} \mathbf{B}_G(C, D) = {} & \frac{1}{8}(\mu_C - \mu_D)^T \left(\frac{\Sigma_C + \Sigma_D}{2}\right)^{-1} (\mu_C - \mu_D) + \\ & + \frac{1}{2} \ln\left(\frac{|0.5(\Sigma_C + \Sigma_D)|}{\sqrt{|\Sigma_C||\Sigma_D|}}\right), \end{aligned} \quad (4)$$

where μ_Z and Σ_Z are the mean vector and covariance matrix estimated for a set Z. $(\cdot)^T$, $|\cdot|$ and $(\cdot)^{-1}$ represent the transpose, determinant and inverse matrix operations, respectively. Using **M(.) = JM(.),** the Stochastic Minimum Distance Classifier, (SMDC), is defined, which associates an unknown region to the closest class in terms of this JM distance.

2.2 Alternatives to the JM Dissimilarity Function.

This section describes alternatives for the JM dissimilarity function developed as a compositions of Stochastic Distances calculated to the individual regions of the classes learning sets and not agregating all pixels, from a determined class, to form just one distribution function.

The first proposed alternative, called the *Stochastic Minimum Mean Distance Dissimilarity Function*, is defined as:

$$\mathbf{M}_{mean}(f_{R_i}, f_{\omega_j}) = \frac{1}{t_j}\sum_{l=1}^{t_j}\mathbf{JM}(f_{R_i}, f_{\omega_j R_l}), \quad (5)$$

where $f_{\omega_j R_l}$ is the probability distribution that models the l^{th} training region assigned to ω_j, which contains t_j training regions in D and f_{R_i} the distribution of the ith unknown region. This function will lead, with equation 1, to the Stochastic Minimum Mean Distance Classifier (SMMDC)

Another alternative is the *Stochastic Nearest Distance Dissimilarity Function*, defined as Equation 6. $\mathbf{M}_{min}(\cdot,\cdot)$ returns the shortest distance between R_i and one of the training regions assigned to ω_j.

$$\mathbf{M}_{min}(f_{R_i}, f_{\omega_j}) = \min\left\{\mathbf{JM}(f_{R_i}, f_{\omega_j R_l}) : l = 1,\ldots,t_j\right\} \quad (6)$$

This function will lead to the Stochastic Nearest Distance Classifier (SNDC). A third alternative is a generalization of Equation (6) which is transformed into a stochastic version of the k*-Nearest-Neighbors* (SkNN) when $\mathbf{M}(\cdot,\cdot)$ is substituted by $\mathbf{M}_{knn}(\cdot,\cdot)$, is defined as:

$$\mathbf{M}_{knn}(f_{R_i}, f_{\omega_j}) = e^{-h_j(f_{R_i})}, \quad (7)$$

where $h_j(f_{R_i}) = \#\left\{(\overline{R}, \omega_j) \in \mathsf{V}_k(R_i)\right\}$, such that $\mathsf{V}_k(R_i)$ is the set of k training regions closest to R_i given a distance $\mathbf{JM}(\cdot,\cdot)$. In other words, $\mathsf{V}_k(R_i) = \left\{(\overline{R}_p, \omega_q) \in \mathsf{D} : 0 < \mathbf{JM}(f_{R_i}, f_{\overline{R}_1}) \leq \right.$ $\left. \leq \cdots \leq \mathbf{JM}(f_{R_i}, f_{\overline{R}_k});\ p = 1,\ldots,k;\ q = 1,\ldots,c\right\}$. In this formalization, $\overline{R}_p$ represents a new indexing of the regions of D based on the proximity to R_i.

Classification of an unknown region is performed associating this region to the least dissimilar class in a similar fashion as Equation 1, where one of those M functions described in this section substitutes M.

2.3. Material and RBC methodology.

The study area is located nearby the Tapajós National Forest, south of Belterra municipality, Pará state, Brazil. A LANDSAT-5 TM image, with 1159×1564 pixels, acquired on July 29, 2010 was used in this research. Fieldwork was conducted during September of 2010 and 115 reference regions, from nine land cover classes of interest, have been acquired, as described in Table 1. Figure 2 illustrates the spatial distribution of the samples of land cover classes in the image. Before classification, this is image is submitted to a segmentation phase to isolate uniform regions that will classified in a subsequent phase. To increase the method efficiency it is important, first, to select properly which TM channels should be used, taking in account the classes of interest, resulting in a two-step process.

A general two-step segmentation phase, is performed as follows:

a) Selecting which bands will be used: intrinsic dimensionality of TM data is in between 3 and 4, based on principal components analysis (PCA) transformation, which means that 4 bands is usually enough to represent the information contained in the 6 channels set of TM data. Using more than 4 channels tend to degrade all image processing transformations due to the curse of dimensionality principle. To improve efficiency, the channels will be chosen in a supervised way, using the feature selection method of Minimum Mean Average JM distance between pairs of classes belonging to the set of classes, conducting to the selection of band numbers of 1, 3, 4 and 5.

b) Applying the segmentor: the chosen set of channels has been segmented by a region growing type of segmentor as implemented in SPRING package (Camara et al., 1996) with similarity threshold of 5 and minimum area of 100 pixels.

Table 1: Summary of the land cover classes

Land cover types	Regions	
	Quantity	Total of pixels
Fallow Agriculture	11	1051
Clean Pasture	18	1317
Clean Pasture with Trees	11	1217
Dirty Pasture	12	1166
Inter. Regeneration	12	1203
New Regeneration	12	1276
Old Regeneration	14	1347
Primary Forest	12	1672
Soybean Agriculture	13	1231
TOTAL	115	11480

3. Results and Discussion

The reference sample regions associated with the classes of interest are used to train the different region-based classification methods. Parameters for SMDC method are estimated collecting non-spatially correlated points of all reference samples for each class. For the other three methods, stochastic distance parameters are calculated for each reference sample and each class and kept apart, using non-spatially correlated points. Figure 3 show the classification results for a detail inside study area. SkNN method was performed with k=3. Table 2 presents the Kappa coefficient of agreement, and its standard deviation, (Congalton e Green, 2009) calculated using randomly chosen 300 reference points for each class, not used in the training phase. Table 2 also presents the p-value when performing a significance test of different means between the best result, which is the SNDC result, and all other results.

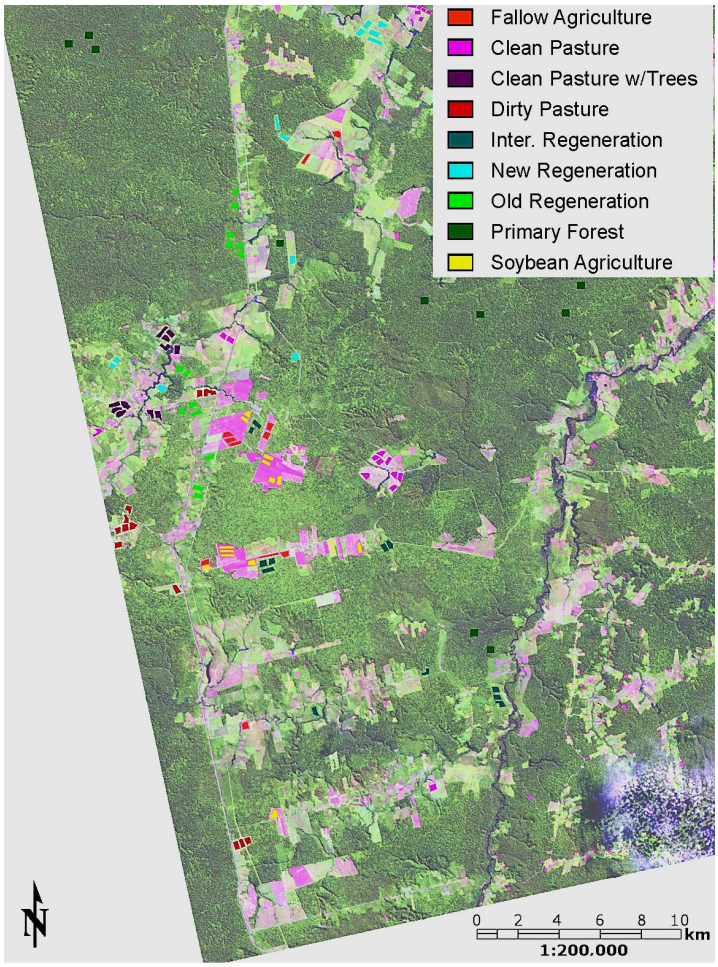

Figure 2: Study image, in (543)RGB composition, and the land cover samples.

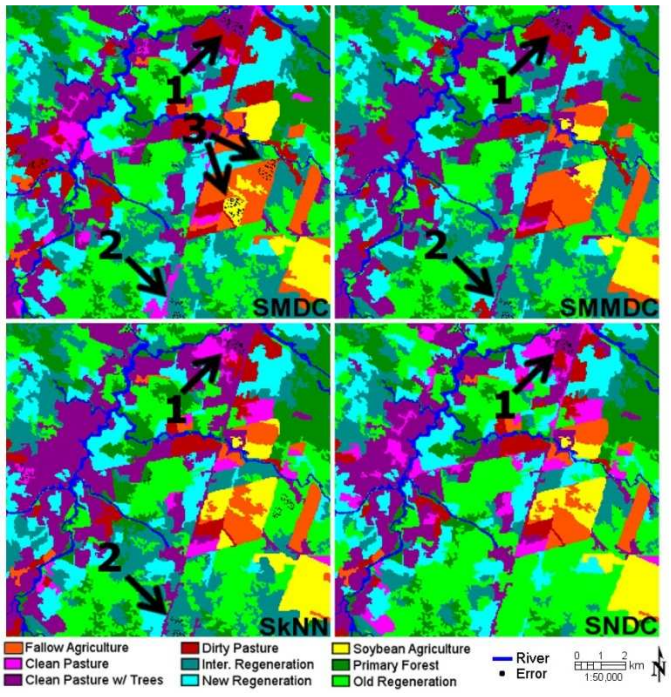

Figure 3: Classification results for a sample area, where black dots represents wrongly classified pixels.

Table 2: Kappa coefficient obtained for each dissimilarity function. Difference of means hypothesis test p-values are relative to SNDC result, which is the best result

	SMDC	SMMDC	SNDC	SkNN
Kappa	0.745	0.730	0.788	0.729
Sd (Kappa)	0.03449	0.03317	0.03876	0.03302
p-value(bi)	0.40	0.254		0.246

Number 1 arrows point misclassified ´clean pasture´ pixels, which have been assigned to ´clean, pasture with trees´ class, in all results. Pixels misclassified as ´primary forest´ (arrow number 2) appear in first 3 results but not in the SNDC result that showed both, in accuracy and visually, as the best result. Arrow number 3 showed some inversion in pixels classification considering the ´fallow agriculture´ and ´soybean agriculture´ class, when using the standard RBC.

Discussion: Nearest neighbor classifiers (NNC) are not normally used in Remote sensing applications because its cost when the number of training patterns is large, which is normally the case when using per point methods. RBC methods, however use regions as a whole conducting to a much smaller and tractable number of training patterns and then turning NNC methods feasible. NNC methods does not supposes unimodality, as most classifiers do, implying in a large usage spectrum. The case pointed by arrow3 supports this interpretation once they represent very dynamic classes; each patch can be in some different cicle stage, which implies in multimodality, which is naturally treated by NNCs methods (SkNN and SNDC). SkNN , with k=3, gave lower accuracy results than SNDC probably because some distribution mode had only one representative.

4. Conclusions.

A study using a LANDSAT-5 TM image was conducted to compare different region based classification methods. The results show that the nearest neighbor type of classifier can outperform standard SMDC classifier. SMMDC method is not recommended by this study. SkNN did not have better results either which is somewhat expected in cases like the one presented in this case, when only about 10 reference patterns are available. This study will continue considering other types of RBC, specially special versions of the SVM classifier.

This type of classifier is not available in any commercial or free software, but its implementation is simple, being a good alternative for RBC methods.

Acknowledgements.

The authors would like to acknowledge the support from CNPq (307.666/2011-5 and 401.528/2012-0) and CAPES (Procad NF2010/486) on the development of this research. Authors also thanks Prof Guaraci Erthal for reviewing this paper.

References

Camara Neto, G.; Souza, R. C. M.; Freitas, U. M.; Garrido, J. SPRING: Integrating remote sensing and GIS by object-oriented data modelling. **Computer and Graphics**, v. 20, n. 3, p. 395-403, May - June 1996. (INPE-6416-PRE/2455).

Congalton, R. G.; Green, K. Assessing the accuracy of remotely sensed data: principles and practices. London, CRC Press, 2008. 2nd ed.,p.183

Liu, Desheng, and Fan Xia. 2010. “Assessing Object-Based Classification: Advantages and Limitations.**” Remote Sensing Letters** 1 (4): 187–94. doi:10.1080/01431161003743173.

Li, Guiying, Dengsheng Lu, Emilio Moran, Luciano Dutra, and Mateus Batistella. 2012. “A Comparative Analysis of ALOS PALSAR L-Band and RADARSAT-2 C-Band Data for Land-Cover Classification in a Tropical Moist Region.” **ISPRS Journal of Photogrammetry and Remote Sensing** 70 (June). International Society for Photogrammetry and Remote Sensing, Inc. (ISPRS): 26–38. doi:10.1016/j.isprsjprs.2012.03.010.

Lu, Dengsheng, Li,G; Moran, E; Dutra, L.V.; Batistella,M.; 2011. “A Comparison of Multisensor Integration Methods for Land Cover Classification in the Brazilian Amazon.” **GIScience Remote Sensing** 48 (3): 345–70. doi:10.2747/1548-1603.48.3.345.

Negri, R. G.; Dutra, L. V.; Sant’Anna, S. J. S. Stochastic approaches of minimum distance method for region based classification. Lecture Notes in Computer Science, Springer Berlin Heidelberg, v. 7441, p. 797–804, 2012.

Richards, J. A.; Jia, X. **Remote Sensing Digital Image Analysis: An Introduction.** 3rd ed., Berlin, Springer, 2005.

Theodoridis, S., e Koutroumbas, K. 2008. Pattern Recognition, Fourth Edition. 4 edition. Burlington, Mass: Academic Press. 2008, p.984.