On the Use of Ancillary Data by Applying the Concepts of the Theory of Evidence to Remote Sensing Digital Image Classification

Rodrigo Lersch [a], Victor Haertel [a,*], Yosio Shimabukuro [b]

[a] Federal University at Rio Grande do Sul ,C.P. 15044 ,Porto Alegre, RS, 91501-970, Brazil – (rodrigo.lersch,victor.haertel)@ufrgs.br
[b] National Institute for Space Research, INPE,C.P. 515,São Jose dos Campos, SP, 12201-970, Brazil yosio@ltid.inpe.br

Key words: Image Classification, Ancillary Data, Theory of Evidence, Uncertainty

ABSTRACT:

This study deals with some applications of the concepts developed by the Theory of Evidence, in remote sensing digital image classification. Data from different sources are used in addition to multispectral image data in order to increase the accuracy of the thematic classification map. Terrain elevation, mean annual temperature , fraction image data of vegetation and shade, NDVI, image texture, as well as probability images estimated from the multispectral image data, are arranged in form of layers in a GIS-like structure. Layers representing *belief* and *plausibility* concerning the labeling of pixels across the image are then derived and later used to help identify mislabeled pixels in the thematic classification map. Preliminary tests using *belief* were performed over an area covered by *natural forest of Araucaria* showing some promising results.

1. INTRODUCTION

In remote sensing supervised image classification, there are a number of distinct factors that may contribute to lower the accuracy of the thematic image. Spectral similarity among classes, non-normality of the within-class data distribution, lack of uniformity in solar irradiation across the scene due to terrain topography are, among others, well known causes for the poor accuracy that sometimes results when using classifiers such as the Gaussian Maximum Likelihood (GML). Different approaches have been proposed by many authors to alleviate each specific cause of low accuracy in the classification process. In this study, we test a method for improving the accuracy in the process of labeling image data by using ancillary data. Different approaches to handle this problem in a quantitative way have been proposed in the literature. A comprehensive overview of the available alternatives to introduce ancillary data into the pixel labeling process can be found in Hutchinson (1982). That author points out three general approaches to combine ancillary data with multispectral image data: before, during or after classification. An example of the before classification procedure is the stratification of the data into smaller areas or strata that are then processed separately. The rationale behind this approach consists in reducing the class variances within each stratum, lessening the likelihood of confusion among classes. A large vegetated area, for example, may be stratified according to the terrain elevation, each stratum comprising more homogeneous vegetated areas showing individually lower variability. Training and classification can then be performed independently on each stratum and the results merged together. Ancillary data can also be incorporated during the classification process. Data can be organized in the form of additional data layers (non-spectral channels) that are then used jointly with the conventional spectral bands in a classifier. Another approach proposed by Strahler (1980) consists of making use of ancillary data to estimate, in a more realistic fashion, the "a priori" class probabilities. In the post-classification approach, areas of confusion, i.e., pixels that have been assigned to classes that are spectrally very similar, are sorted individually and assigned to the most likely class by making use of information that can help discriminating distinct classes, such as digital elevation models data (DEM), slope or aspect. The decision rules in this case are usually deterministic in nature. These approaches imply the use of multisource data, sometimes including qualitative data, i.e., variables that originally are not in numerical form. Bruzzone et al. (1997) investigated the use of Landsat-TM image data along with texture and ancillary data (terrain elevation, slope and aspect) in the classification of spectrally complex areas (rural areas), using both parametric (statistical) e non-parametric (neural networks) classifiers. In the statistical approach, a total of 17 features were introduced, including both multispectral image data and texture data. In this case, the ancillary data (terrain elevation, slope and aspect) were used to improve the estimates of the land-cover classes *a priori* probabilities. The evidential approach expressed in the Dempster-Shafer theory of evidence is investigated by Lee et al. (1987). In addition to a comprehensive review of the basic concepts in this theory, the authors explore statistical and a evidential methods for combining multi-source remote sensing image data and spatial data in general. They tested both methods using Landsat MSS image data, dealing with the visible and infrared bands as two separate data sources. The results reported in their work, show that the statistical approach performed better than the evidential method. It should be noted, however, that the data used in the experiment can be reasonably modeled by the multivariate Gaussian distribution. It does not take advantage of one main advantages of the evidential method, i.e., of the fact that it is not distribution based. Evidential methods were also explored by Moon (1990) to integrate geological and geophysical data. Gong (1996) compares evidential reasoning with neural networks methods in the classification of multisource spatial data. The author makes use of multisource data set which includes Landsat-TM image data, aeromagnetic, radiometric and gravity data for geological mapping purposes. This approach allows the introduction of data layers such as the ones commonly used in GIS into the classification process, along with multispectral image

* Corresponding author.

data. In addition, the theory of evidence also allows the introduction of the concept of uncertainty or ignorance into the classification process. Uncertainty or ignorance in this case can be caused by incomplete data or lack of sufficient information. The aim in this study is to investigate the use of the Theory of Evidence for vegetation cover classification purposes. Information conveyed by auxiliary data is used to generate layers of *belief* and *plausibility* as proposed by Dampster-Shafer. These layers may be later used to detect mislabeled pixels in the thematic image produced by a classifier implementing the available multispectral image data set, such as the GML classifier. In this study a procedure implementing the information conveyed by *belief* is tested in the identification of areas covered by *natural forest of Araucaria*. *Araucaria angustifolia* is a native conifer species that occurs in association with other gymnosperma species, the *Podocarpus lambertii*, as well as several trees or shrubs such as *Slonea monosperma*, *Symplocus uniflora, Ocotea pulchella, Rapanea venosa, Feijoa sellowiana, Eugenia opaca* and epiphytes such as *Polypodium squamulosum* and *Tillandsia tenuifolia* (Ferri, 1980). The areas covered by *natural forest of Araucaria* are restricted to small areas in southern Brazil and have proved to be particularly difficult to be accurately identified by means of multispectral image data only, such as Landsat TM data. To alleviate this problem and improve the classification accuracy, the principles provided by the theory of evidence are here applied. In addition to Landsat multispectral data, we have also investigated other possible sources of data that may bear discriminant power with respect to the label *natural forest of Araucaria*. A total of 15 auxiliary variables were tested, namely digital elevation data, mean annual temperature and precipitation, soil classes, NDVI, fraction data of shade and vegetation, and 8 additional variables associated with image texture. An additional layer conveying the joint contribution of the multispectral bands was also added to this set. In a preliminary step, we investigated the relative importance of each individual variable, i.e., its discriminant power with respect to the label *natural forest of Araucaria*, in the context of the scene being analyzed. The information conveyed by these 16 variables was initially expressed in terms of evidence in support of the label *natural forest of Araucaria*, for every pixel in the image.

The evidence in support of the label *natural forest of Araucaria* was estimated for each individual variable. For the variables presenting a normal distribution, the evidence was estimated by:

$$P_a(X) = \frac{G_a(X)}{\sum_{i=1}^{m} G_i(X)} \qquad (1)$$

where $G_i(\boldsymbol{X})$ represents the GML decision function associated with class ω_i , estimated in each case for the variable under consideration only; *a* indicates the class *natural forest of Araucaria* and *m* the number of spectral classes present in the scene. In the case of the variables that are not normally distributed, membership functions and expert knowledge were used.

The layers expressing evidential data were then structured in a GIS-like format, in layers spatially registered and within the interval [0,1], with 0 meaning no evidence supporting the label *natural forest of Araucaria* and 1 indicating full evidence available. In this context, an additional layer bearing the evidential information conveyed by the spectral data is added to the set. This set of evidential information layers were spatially registered, geocoded and then combined according to the Dempster-Shafer approach to produce the layers of *belief* and *plausibility*.

2. THE DAMPSTER-SHAFER THEORY OF EVIDENCE

In this section some basic concepts of the theory of evidence are reviewed. In the context of this study, the two most relevant characteristics of the Theory of Evidence are: (1) the introduction of multi-source data, each data source being treated separately in a way that does not require all data being originally in numerical form, and (2) implementing the concept of uncertainty or ignorance (Lee et al., 1987, Richards and Jia, 1999, Eastman, 1999).

One feature that distinguishes the theory of evidence from other commonly used methods implementing multi-source data relates to the data form. Most of the methods implementing multi-source data require that the data must be in numerical form. The theory of evidence allows, however, different data sources to be treated separately, not requiring all variables to be originally in numerical form. In this case, the translation of thematic data into numerical form is performed by the analyst by assigning to each variable the amount of evidence in support of an hypothesis, (in the present case the label *natural forest of Araucaria*), producing in this way layers of evidence. Whenever the data is originally in numerical form (such as terrain elevation, for instance), the task of estimating the amount of evidence supporting the hypothesis can be more easily done by making use of probability functions as in (1) or membership functions. The amount of evidence supporting a given hypothesis normally ranges in the interval [0,1]. If, however, a certain degree of uncertainty or ignorance with respect to the pixel's label is present, then this fact can be taken into consideration by restricting the upper limit of this range (maximum evidence) to a lower value, say, 0.8, leaving the remaining evidence (20 %) as ignorance or uncertainty. The Theory of Evidence makes use of the following concepts: *mass of evidence* also known as *basic probability assignment*, *belief* known by some authors as *support*, *plausibility* and *belief interval,* also known as *evidential interval* (Richard and Jia, 1999, Lee et al., 1987, Gong, 1996). The *mass of evidence* or *basic probability m(X)* estimates the evidence available in support of a given hypothesis. These concepts can be better illustrated through an example. Consider a classification problem consisting of four classes ω_1, ω_2 , ω_3 and ω_4. In addition to being exclusive, assume that this set is exhaustive, which means that in practice one of the classes could be referred to as "other", (Lee et al., 1987). Let us assume that, for a pixel under consideration, the likelihood for the four classes can be estimated to be proportional to 2:2:1:1. Further, assume that a degree of uncertainty exists, due to the quality of the data or to the classification procedure. The analyst may estimate that under the present circumstances his level of confidence in the labeling of this pixel is about 90%, leaving the remainder 10% as uncertainty or ignorance. Therefore, amount of evidence available can be expressed by the *mass of evidence* (or the *basic probability assignment*) -*m*- as follows:

$$m(\omega_1)=0.3 \quad m(\omega_2)=0.3 \quad m(\omega_3)=0.15 \quad m(\omega_4)=0.15$$
$$m(U)=0.1$$

where U stands for the uncertainty in the labeling process. Obviously, these basic probabilities should add to one and may be derived in an empirical way from the knowledge the analyst has about the nature of any particular situation. This concept can be further generalized by assuming that class ω_3 actually represents the union of classes ω_1 and ω_2 ($\omega_1 \cup \omega_2$), i.e., the information available allows the analyst to allocate, for the case of a pixel under consideration, a mass of evidence $m(\omega_3)$ of 0.15, meaning

that this is the degree of evidence available, indicating that the pixel belongs to either ω_1 or ω_2 , although it can not be said to which one.

Different ways to estimate the mass of evidence can be seen in the literature. Gong (1996) makes reference to two basic ways of training an evidence based algorithm, i.e., of estimating the mass of evidence: the occurrence-frequency tables and the normal distribution model (similarly to the maximum likelihood classifier). Lee et al. (1987) proposes the following procedure to assign the mass of evidence to each particular pixel: first estimate the uncertainty in the labeling process and second, assign the remaining mass of evidence based on the relative likelihood of each label.

If two independent sources of information (a, b) are available, the combined mass of evidence $m_{a,b}(z)$ can be obtained by means of the *orthogonal sum*:

$$m_{a,b}(z) = \frac{\sum_{x \cap y = z} m_a(x) \cdot m_b(y)}{\sum_{x \cap y \neq \varnothing} m_a(x) \cdot m_b(y)} \qquad (2)$$

where $\varnothing$ represents the null set. The orthogonal sum (2) can be applied repetitively whenever more than two sources are present. The *belief* or *support* (represented by *bel*) for a given hypothesis represents the total support for an hypothesis. It can be estimated by adding the masses of evidence supporting the hypothesis, i.e., the label under consideration (X) and any of its subsets (Y). In a more concise notation:

$$bel(X) = \sum m(Y) \qquad \forall\, Y \subseteq X$$

In the above example, the *beliefs* are:

$$bel\,(\omega_1) = 0.3, \quad bel\,(\omega_2) = 0.3, \quad bel\,(\omega_3) = 0.75$$

The *plausibility* denoted by *plau* of an hypothesis is defined as the complement of the total support for contradictory hypothesis:

$$plau(X) = 1 - bel(NOT(X))$$

In the case of the previous example, the *plausibilities* for ω_1, ω_2 and ω_4 are:

$$plau\,(\omega_1) = 0.55, \quad plau\,(\omega_2) = 0.55, \quad plau\,(\omega_4) = 0.25$$

A straightforward interpretation can be assigned to these two concepts: *belief* represents the firm evidence available in support of a hypothesis, whereas *plausibility* represents the maximum amount of evidence supporting the hypothesis, i.e., the degree to which the hypothesis seems to be correct or, in other words, the evidence favoring the hypothesis even considering that firm data may be missing. *Belief* and *plausibility* can therefore be understood as the minimum and the maximum amount of available evidence in favor of a hypothesis or, in this case, a particular pixel labeling.

The *belief interval* or *evidential interval* is defined as the difference between the plausibility (upper bound) and *belief* (lower bound) and can be seen as the uncertainty in accepting or rejecting the hypothesis.

The approach provided by the Theory of Evidence may be helpful for image classification problems in remote sensing applications, in cases where no accurate conclusions can be drawn from the available multispectral image data only.

3. CASE STUDY

The methodology provided by the Theory of Evidence was tested in this study, in the identification of areas covered by *natural forest of Araucaria*. A test area in southern Brazil was used for testing purposes. The required computer programs were produced in MATLAB environment.

The implementation of this methodology implies three steps: (1) the identification of the variables to be used in this problem, i.e., variables showing discriminating power with respect to the label *natural forest of Araucaria*, (2) a criterion to estimate the associated degree of uncertainty, and (3) a criterion to estimate the amount of evidence provided by each individual variable. Initially a survey of the variables which are thought to bear a high degree of correlation with the label *natural forest of Araucaria* was performed. In addition to multispectral image data, the following variables were initially investigated: terrain elevation above sea level, mean annual precipitation, mean annual temperature, category of soil, NDVI, fraction data of vegetation and shade, and image texture estimated by the following eight variables: contrast, correlation, dissimilarity, entropy, homogeneity, mean value, angular momentum and standard deviation. These data, obtained from different sources, were arranged in layers, spatially registered and geocoded in a GIS format.

The next step consisted in expressing these layers in terms of *mass of evidence* supporting the hypothesis *natural forest of Araucaria*. Two criteria are required at this stage, one to estimate the amount of evidence and another to estimate the amount of uncertainty involved. It is generally more convenient to approach this problem by estimating the uncertainty. In this study we follow the approach proposed by Lee et al. (1987), estimating the uncertainty associated with each individual variable by the classification error in the corresponding thematic map. The accuracy in this case refers to the user's accuracy, which estimates the probability of any given pixel in the thematic image being correctly classified (Congalton, 1991).

Variable	User's Accuracy
Multi-spectral data	0.897
Terrain elevation	0.510
Mean annual temperature	0.515
Mean annual precipitation	Negligible
Soil categories	Negligible
NDVI	0.463
Fraction of shade	0.513
Fraction of vegetation	0.352
Contrast	0.338
Correlation	0.377
Dissimilarity	0.372
Entropy	0.413
Homogeneity	0.454
Mean value	0.613
Angular momentum	Negligible
Standard deviation	0.360

Table 1. User's accuracy regarding the label *natural forest of Araucaria*, yielded by each individual variable

To this end, samples of the 21 spectral classes identified on the scene were collected for training and testing purposes. Whenever the data presented a normal distribution, the Gaussian Maximum Likelihood classifier was applied. For the remaining cases (terrain elevation data, mean annual temperature and precipitation and soil categories) the Minimum Euclidean Distance classifier was employed. The estimated user's accuracy is shown in Table 1. Information regarding steps (1) and (2) can thus be obtained from data listed on this table. Table 1 suggests that three variables bear no discriminating power with regard to the problem (mean annual precipitation, soil categories and texture's angular momentum), leaving 13 variables to be used. The uncertainty associated with each individual variable can thus be estimated (1-user's accuracy). Step (3) consists in estimating the mass of evidence for each of the selected variables. For the completion of this task, two different approaches were pursued. For the normally distributed variables, equation (1) was applied. Otherwise, the concept of membership function associated with expert knowledge was employed. In the first case, the mass of evidence supporting the label *natural forest of Araucaria* is estimated for each individual variable, by re-scaling the values produced by equation (1) to within the interval [0, (1-uncertainty)]. For every pixel across the scene, this re-scaled value is assumed to estimate the mass of evidence supporting the label *natural forest of Araucaria*. The mass of evidence supporting the label *others* is, for each pixel, equal to one minus the sum of the uncertainty plus the mass of evidence supporting the label *natural forest of Araucaria*, i.e., the two masses of evidence plus the uncertainty should add to one.

A similar approach was applied to the variables that are not normally distributed (terrain elevation and mean annual temperature). In this case, however, expert knowledge modeled by a membership function was employed to estimate the mass of evidence supporting the label *natural forest of Araucaria* and the label *others*. It is known that *natural forest of Araucaria* does not occur in areas below the elevation of 500 meters above sea level, and the likelihood of it occurring increases as the terrain elevation also increases. Based on this knowledge, a digital elevation model was initially produced for the study area. Next, a membership function was selected to relate terrain elevation with the corresponding *mass of evidence* in support of the label *natural forest of Araucaria*, in a pixelwise fashion. A sigmoid function was deemed adequate to represent this association. The first control point that defines the particular shape of the sigmoid, i.e., the terrain elevation at which the mass of evidence starts rising from zero was set at 500 meters above sea level. The second control point at which the available evidence reaches its maximum value (1-uncertainty), was set at the highest elevation occurring in the area (1,300 meters). It is also known that natural occurrences of *natural forest of Araucaria* have been found only in regions where the mean annual temperature is lower than 18 °C. Based on this knowledge, a decreasing sigmoid membership function was selected to estimate the *mass of evidence* associated with the mean annual temperature. The control points in this case were set at 18 °C (*mass of evidence* equal to zero) and 14.5 °C, the lowest mean annual temperature in the region (*mass of evidence* equal to 1-uncertainty). As in the previous cases, the evidence supporting the label *other* was estimated according to the condition stated above, i.e., the uncertainty and the masses of evidence should add to one.

The mass of evidence for the labels *natural forest of Araucaria* and *others* plus the uncertainty can then be combined in a pixelwise fashion to produce the combined mass of evidence for the label *natural forest of Araucaria*. To this end, the algorithm known as the orthogonal sum (2) was used. Thus, layers displaying *belief* and *plausibility* associated with the label *natural forest of Araucaria* can be produced. As mentioned earlier, *belief* estimates the solid evidence supporting the hypothesis, i.e., the available evidence that a pixel does belong to the class *natural forest of Araucaria*, whereas *plausibility* is the maximum available evidence supporting the hypothesis. A possible use for the information conveyed by *belief* and *plausibility* consists in detecting errors of commission and omission respectively. In this study it is investigated the use of the information contained in the layer *belief*. To this end, a thematic image was produced by applying the MGL classifier to the 6 reflective Landsat-TM spectral bands. The resulting thematic image shows the 21 land-cover classes present on the test area. For the purposes of this study, the thematic image was re-classified into two classes, namely *natural forest of Araucaria* and the remaining 20 classes, grouped together under the label *other*. The investigation was performed using a segment (922 rows-1947 columns) in a Landsat-TM scene, for which ground truth data was available. The availability of reliable ground truth data allowed the estimation of the user's accuracy for the label *natural forest of Araucaria* as equal to 0.7209. Comparing the thematic image data with the ground truth data, a number of mislabeled pixels can be detected, resulting in both errors of commission and omission. As expected, pixels wrongly labeled as *natural forest of Araucaria* were associated with low values for *belief*. Removing pixels labeled *natural forest of Araucaria* but showing low degree in *belief* is an obvious choice to improve the user's accuracy. The question at this point, however, is how to estimate a suitable value for the threshold to be applied to the layer *belief*. Obviously, the higher the selected value for the threshold, the lower the likelihood of including errors of commission in the label *natural forest of Araucaria*. Setting the threshold at values which are excessively high, however, increase the likelihood of removing pixels correctly labeled as *natural forest of Araucaria*, i.e., introducing errors of omission in the thematic image, decreasing therefore the producer's accuracy. This fact is illustrated in Figure 1, which shows the user's accuracy, the producer's accuracy and the mean accuracy as a function of the threshold applied to *belief*. As expected, low values for the threshold result in a large number of pixels mislabeled as *natural forest of Araucaria* (errors of commission) which translates into lower values for the user's accuracy. The number of pixels mislabeled as *natural forest of Araucaria* decreases as the threshold increases, resulting in a correspondent increase in user's accuracy for the label *natural forest of Araucaria*. This approach, however, is not reasonable as the number of errors of omission starts increasing, leading to an increasingly lower value for the producer's accuracy.

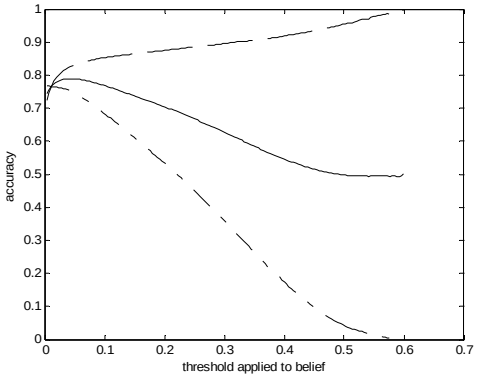

Figure 1 Accuracy as a function of threshold applied to *belief*. Dashed line: user's accuracy; dash-dot line: producer's accuracy; solid line: mean accuracy

Data in Figure 1 can be used by the analyst for a final decision about the threshold that better suits his/her needs. In this study we propose a threshold that maximizes the mean accuracy in the thematic image, which in this particular experiment lies around 0.0431 resulting in a user's accuracy equal to 0.8264, a producer's accuracy equal to 0.7519 and a mean accuracy equal to 0.7891. The corresponding estimated values in the thematic image produced by the GML classifier is 0.7230, 0.7665 and 0.7448 respectively. A significant increase in the user's accuracy and in the mean accuracy has been obtained, while a small decrease in the producer's accuracy occurred.

4. CONCLUSIONS

In this study we present and test a methodology focused on the use of ancillary data to help improve the accuracy in remote sensing digital image classification. A new approach aiming at the use of the concept of *belief* to increase the accuracy of the thematic image is proposed and tested. The proposed methodology is deemed especially useful whenever we are dealing with classes that are spectrally difficult to be identified accurately. The adequacy of this methodology was assessed by using data available from a test area in southern Brazil, covered by a forest type known as *natural forest of Araucaria*. This class is known to be difficult to be accurately classified. Tests were performed using Landsat TM multispectral image data and additional ancillary data available. Encouraging results were obtained suggesting that further work would be fruitful. It is especially suggested further research work with regard to the use of *plausibility* data for additional improvement concerning the accuracy of the thematic image.

References

Bruzzone, L., Conese, C., Maselli F., and Roli, F.,1997. Multisource classification of complex rural areas by statistical and neural network approaches. *Photogrammetric Engineering and Remote Sensing*, 63(5), pp. 523-533.

Congalton, R., 1991. A review of assessing the accuracy of classification of remotely sensed data. *Remote Sensing of Environment*, 37, pp. 35-46

Eastman, J. R., 1999. *Guide to GIS and Image Processing*, vol. 2, Clark Labs, Clark University.

Ferri, M.G., 1980. *Vegetação Brasileira*, University of São Paulo Press, São Paulo.

Gong, P., 1996. Integrated analysis of spatial data from multiple sources: using evidential reasoning and artificial neural network techniques for geological mapping. *Photogrammetric Engineering & Remote Sensing*, 62(5), pp. 513-523.

Hutchinson, C.F., 1982. Techniques for combining Landsat and ancillary data for digital classification improvement. *Photogrammetric Engineering and Remote Sensing*, 48(1), pp. 123-130.

Lee, T., Richards J. A., and Swain, P. H., 1987. Probabilistic and evidential approaches for multisource data analysis. *IEEE Transactions on Geoscience and Remote Sensing*, 25(3), pp. 283-293.

Moon, W. M., 1990). Integration of geophysical and geological data using evidential belief function, *IEEE Transactions on Geoscience and Remote Sensing*, 28(4), pp. 711-720.

Richards, J. A. and Jia, X., 1999. *Remote Sensing Digital Image Analysis, an Introduction*, 3rd edition, Springer, Berlin.