Amazonia Landscape Mapping and Biodiversity Estimation

Luigi Fabbro

Associação Amazônia, Manaus (AM) Brazil
http://amazonia.org

Associazione Amazonia ONLUS, San Polino , 53024 MONTALCINO (SI). Italia
luigi.fabbro@amazonia.org

,

Abstract
This project aims to set-up and test a methodology for landscape mapping of Amazonia to assess the biological diversity by applying the knowledge and know-how of the local communities and involving them in Remote Sensing and Geographic Information System technology. The project will provide ground truth for the remote sensing of Amazonia and contribute to the baseline data sets of pristine biotopes against which "trouble spots" can be monitored.
The concrete output of this project will be a GIS mapping the landscape of the Lower Rio Jauaperì area and in particular to map the presence of selected taxa belonging to the local community taxonomy. These maps may be used to determine biodiversity indexes like the alpha, beta, gamma and other measures of biodiversity so that comparable data is produced (Fabbro, 2000).

BIODIVERSITY REDUCTION

In a recent issue of Nature May 11, 2000 dedicated to biodiversity the following estimates were given:

- 5% to 20% of the species have already been lost
- tropical forest is lost at a rate of 0.5% to 2% a year
- 50% to 70% of species are found in tropical forests
- loss of 50% to 80% of thc species of an ecosystems causes the collapse of most biogeochemical ecosystem processes

Only l0 percent of the existing species are known and classified. We know even less about their distribution and still less about their interaction.
Indeed, " like children playing with fire, we do not fully understand, and therefore cannot predict, the ultimate consequences of tampering with global biodiversity" .

SUSTAINABLE DEVELOPMENT

In addition to testing new methodologies for biodiversity assessment, this project aims to contribute to:

- Environmental education focusing on the economic benefits of protection and sustainable utilization of biodiversity
- Establishing Leapfrog Technology (Computers, GIS, GPS, Internet) in Amazonia
- Creating local community friendly computer interfaces
- Capacity building for bio-prospecting, within the framework of the "Extractive Reserves" concept developed by Chico Mendes and Mary Allegretti .

The project is an exercise in sustainable development, our best hope yet to stop deforestation and avert the biological catastrophe. It aims to create new jobs and raise living conditions of local communities encouraging them to remain in the forest and protect it.

STUDY AREA

Xixuaù-Xiparinà Nature Reserve and Lower Rio Jauperi basin , State of Roraima, Brazil.
0° North - 2° South 61° West - 62° East (110 Km x 220 Km). The unexplored, pristine rainforest harbors a rich wildlife including several threatened mammal species. Being placed between hydrographic basins of different geological ages, it is of great biological relevance. The total population is 570 people in 5 communities. The ***Associação Amazônia***, largely composed of people belonging to the local communities, has been working in the region since 1991.

BIODIVERSITY ASSESSMENT

Biodiversity is the variety of living organisms considered at all levels of organization, from genetics through species, to higher taxonomic levels, including the variety of habitats and ecosystems, as well as the processes occurring therein
The "ecosystem approach" is the primary framework for the implementation of the Convention on Biodiversity and requires that in the assessment of biodiversity:

- all the components of biodiversity should be considered .
- all forms of relevant information, including scientific and local knowledge, innovations and practices should be considered
- identification and monitoring of ecosystems and habitats as well as identification, monitoring and assessment of species should be included

Local community knowledge is used in the biodiversity assessment. The Convention on Biological Diversity requires the application of the knowledge of local communities. The Subsidiary Bodies for Scientific, Technical and Technological Advice (SBSTTA) said that traditional and indigenous people have the knowledge and perspectives that should be included in current taxonomic systems. Scientific taxonomy covers more species than indigenous taxonomy, yet it covers only 10% of existing species.
Fundamental to the present project) are the views expressed in the report of the 2nd meeting of the SBSTTA :

> ***"The challenge is not to find the ways to integrate, in modern management practices, knowledge, innovations and practices of indigenous and local communities. Rather, it is to define, in collaboration with indigenous and local communities, which modern tools may be of help to them, and how these tools might be used, to strengthen and develop their own strategy for conservation and sustainable use of biological diversity, fully respecting their intellectual and cultural integrity and their own vision of development."*** (SBSTTA, 1996a).

The SBSTTA has listed and summarised several existing methodologies for the assessment of Biodiversity (SBSTTA, 1996b).. These generally use remotely sensed data, pre-existing cartography and inventories, GIS systems , and indicator groups. The species inventoried are commonly vascular plants and vertebrates.
The Rapid Biodiversity Assessment (MacQuarie University) uses locally functional schemes for classification and identification as an alternative to formal and correct species identification by expert taxonomist [8]
The National Conservation Review (Shri Lanka Forest Department) utilizes gradient-directed sampling. Transects are selected deliberately to transverse the steepest environmental gradients present in the area, while taking into account access routes This technique is considered appropriate

for rapidly assessing species diversity in natural forests, while minimising costs, since gradient transects capture more biological information than randomly placed transects of similar length.

GEOGRAPHIC INFORMATION SYSTEM (GIS)

The GIS utilized, running under Linux, is based on GRASS . It has the following components: on-site workstation, off site workstations, web server and. java implemented web client. GRASS was selected because it is particularly suitable for managing parks and forests and generally large areas. GRASS is freely available, distributed under GNU General Public License, the complete source code is provided allowing modifications of the programs to fulfill special needs, it is maintained and developed by programmers world wide and has a large number of users.

Other programs, all freely available are also used:

WEKA: The Waikato Environment for Knowledge Analysis is a collection of machine learning algorithm for solving real-world data mining problems. It was developed at the University of Waikato, Hamilton, New Zealand. It is freely available and is distributed under GNU General Public License. It has implemented several schemes of classification, like decision trees inducers, naive Bayes, multi-layered perceptron as well as several schemes of numeric prediction and "meta-schemes" like bagging and boosting.

6s is a freely distributed program developed by NASA and used for atmospheric and terrain correction

Xgobi is system for multivariate data visualization. It has been developed at AT&T Labs and Iowa University and is freely available. It is possible to visualize Tasseled Cap Transformations and Parallel Coordinates Representation to be used in Spectral Mixture Analysis.

gstat is a program for the modelling, prediction and simulation of geostatistical data in one, two or three dimensions. It affords variogram modeling and different manners of kriging. It is freely distributed under the GNU General Public License.

Other programs are being developed internally like the Java GIS Client and the Spectral Mixture Analysis program.

DATA SOURCES

Existing Cartography and Digital Maps

The following cartography (duly registered and rectified) is used: IBGE Topographic maps 1:250.000 and *Levantamento Planinimetrico do Xixuaú* 1:50.000 as well as the following digital maps: DCW, GT003.0, HYDRO 1 K, GDEM,

Remotely sensed data

Time series spectral data from Landsat TM-5, Landsat TM-7 , Envisat MERIS, as well as SAR data from ERS-1 ERS-2 Envisat and JERS-1 are used.

The satellite imagery is registered and rectified the expected accuracy being 10 meters. Atmospheric correction is made using the freely available program 6s. The shadows caused by geographic features are cancelled with the help of DEM data. obtained from IBGE 250:000 topographic maps and from GT003 and GDEM digital maps Clouds and clouds shades are masked

Higher spectral resolution provides better classification of the land cover and thus full use must be made of the new instruments with higher spectral resolution such as Envisat MERIS. These however have low spatial resolution and thus should be fused with the higher spatial resolution of LANDSAT yet conserve the spectral content. Landsat 7 also has a panchromatic band at 15 meters resolution that could be used to sharpen the low spatial resolution data and the thermal data.

The fusion of the data from different spectral and space resolution of MERIS and LANDSAT including the latter thermal band and panchromatic band is performed using the Multisensor Multiresolution Technique (MMT) (Zhukov et al., 1999) (Minghelli-Roman et al., 2001).
In this way all spectral data can be sharpened to the 15 meters spatial resolution of the LANDSAT panchromatic band .

Seasonal variation of spectral data reflects phenological features of vegetation and thus can be used for vegetation type classification.
The spectral data obtained on different seasons can be composited to obtain a single multispectral image. For example two MERIS images taken on different dates consisting of 14 bands produce a single image consisting of 28 spectral bands

Principal Component Analysis has traditionally been employed in Remote Sensing as a data reduction and de-correlation technique. Principal components are de-correlated (independent to the second order) and thus become, for example, available to Bayesian methods which require the data to be independent (to the first order).
The first component contains most of the variance while the last components contain mainly noise. In many landscapes, for example, the seven bands of Landsat TM data can be shown to have an inherent information content of about 2-3 components. In the case of temporal data most of the variance due to seasonality is present on the second and higher components.
The images resulting from PCA analysis are however of difficult interpretation. On the other hand NDVI and other Vegetation indexes like Tasseled Cap Transformation (Vegetation Greenness, Vegetation Wetness and Soil Brightness) as well as endmembers resulting from spectral unmixing provide reduced data which can easily be interpreted.
Spectral Mixture Analysis (SMA) assumes that the reflectance of each pixel is a linear combination of contributing sub-pixel components. The spectral signature of these components, or. endmembers can be obtained from the image itself via Principal Component Analysis (PCA) and Parallel Coordinates Representation (PCR). For this purpose, the freely available graphical analysis program XOBI is utilized. After rejection of the highest and most noisy components Landsat can provide 3 or 4 endmembers and MERIS 10 or 12 .

Local community
Higher taxon richness indicators are useful surrogates for biodiversity (Williams et al., 1994) (Williams , Gaston 1994) (Williams et al., 1997). This permits the performance of biodiversity surveys based on indigenous taxonomy. Local community biodiversity surveys can be performed at much lower costs than scientific surveys but are nevertheless capable of providing comparable results.
Information from the local communities is collected in individual as well as group sessions, by previously trained local community members. Local community names of animal, plants, habitats, as well as geo-bio-physical categories are collected and registered.
The correspondences between the indigenous taxonomy and scientific taxonomy is found and thus access to scientific knowledge is enabled. Both local names and scientific names are used to search Internet. The taxa is then categorized according to their economic function (food, pests, etc) , ecological function (keystone species, etc), the interaction with other species , habitat class, territory size, etc. The taxa are also localized on the food web and in the tree of life (philogeny).
This information and the information directly availed from local experts is utilized to obtain joint probabilities for the Bayesian predictors.

Local experts also provide information about presence of taxa at given locations For this purpose a topographic map of the area on a 1:25.000 scale , the scale usually employed in landscape ecology, is prepared beforehand using existing topographic maps on a 1:250.000 scale and satellite

imagery. Using the MMT technique Landsat and Meris data can be sharpened to 10 meters resolution, thereby making possible the generation 1:25.000 topographic maps.
Different views of the area are also provided including False Colors, NDVI, Tasseled Cap Transformation Vegetation Greenness, Vegetation Wetness and Soil Brightness as well as Endmembers as an aid to local experts to indicate on the map the presence of the various taxa and various habitats. Humans are able to detect and recognize as many as 10.000 distinct objects (Biederman, 1985) under varying viewing conditions, while state-of-the-art object recognition system can recognize relatively few objects (Shah, S. and Aggarwal, J. K. 1999b)
The information is written on top of transparent plastic sheets overlaid on top of large maps produced by collating size A4 printouts. Successively the information is entered into the GIS; After having become more familiar with computers and GIS technology local experts will enter the data directly into the GIS using r.digit
As there are many experts available and the data samples can be very large, statistical methods are applied to obtain expectations, confidences and other statistics.

Ground Truth
Expeditions are organized to selected locations to obtain ground truth. These expedition are integrated in the eco-tourism program of the Amazonia Association and provide part of the funding necessary to finance this project. Random sampling, systematic sampling as well as preferential sampling are performed. Systematic sampling is used for kriging interpolations. Preferential sampling is used for the localization of gradient transects as well as for resolving conflict among predictors and for validation.
The ground truth data is used to validate the predictions made and provide a measure of the confidence of the predictions and is also used as training sets.

Ground truth data is collected using two techniques.

1. A list of the taxa belonging to the local community taxonomy are prepared and are used as a check list of the taxa present in the transects. Transects are precisely geo-located and exhaustively searched.
2. During the journey to the transect the description of the encountered habitat, and the presence of taxa are registered on a minidisk recorder time-synchronized with a GPS thus allowing the geo-location of the information recorded which is then transcribed, opportunely coded, geo located and inserted into the Data Base. The data can then be inserted as categories into raster maps of the GIS

TAXA PRESENCE PREDICTION
The available data from remote sensing, local community , Internet, existing maps and ground truth are fused to predict, employing a Bayesian framework, the presence of the various taxa at any location in the area.

Features
From the available data various features are extracted and a selection performed to eliminate redundant, irrelevant and noisy features.
LANDSAT and MERIS time series data are fused with MMT and the data sets of different dates are composited. The resulting hypespectral data is reduced to fewer, de-correlated and less noisy band layers via PCA. Different views of the remotely sensed spectral data can also be generated like NDVI, Tasseled Cap Transformation, endmembers. The original spectral data can also be utilized directly and the vegetation characteristic spectral features extracted (Qian Tan et al., 2000).

Another set of rasters are derived from SAR data: images from different dates and based on different polarizations are again composited and then reduced to fewer components and de-correlated via PCA.
Another set of rasters are generated containing information about texture and spatial relations. For this purpose the command i.texture is used for 3x3 neighborhoods and r.mcalc for larger neighborhoods
The locations defined by taxa presence either from ground truth survey or from local experts indications can be used to provide, through i.gensigset , spectral signatures. These can then be employed by i.smap to extrapolate the prediction of the taxon presence to unexplored and unknown areas.
The patches defined by taxa presence are measured using the GRASS landscape ecology programs r.le to provide various metrics including diversity metrics like Shannon and Simpson indices and new rasters are generated using these measurements as categories.

Multiple Scales
The pattern detected in any ecological mosaic is a function of scale and the ecological concept of spatial scale encompasses both extent and resolution. Scale should be defined from the perspective of the organism or ecological phenomenon under consideration and therefore it is necessary to understand scale variation. Methods available to investigate the scaling properties of remotely sensed data include multiscale variance, local variance variogram analysis fractal dimensions, spectral analysis and wavelet analysis.
The results of the investigation and ecological considerations provides the characteristic scale of the landscape for each taxon and helps reveal the often self-similar nested patching of the forest.
The first layer of patching is obtained by partitioning the landscape into : water areas, *terra firma*, *igapò*, and various categories of vegetation cover as well as the border areas. These areas are selected using the classification of the vegetation cover in the IBGE topographic maps of the area in scale 1:250000, supervised clustering, and region growing.

Feature Selection
The performance of learning predictors can be improved through feature selection. Different algorithms for feature selection, based on different criteria, have been proposed and most of these are present in the program WEKA.
The set of features is partitioned into effective subsets and these are used by different predictors.
The set of features employed by each individual predictor is selected so that the individual predictors are both accurate and make their errors on different parts of the input space (Hall, Holmes, 2000).

Discretization
Most classification tasks involve learning to distinguish among nominal class values and if the attributes are continuous values these must be transformed into nominal values or discretized.
Several discretization methods have been proposed. The simplest is equal interval width which divides the continuous feature into a user selected number of bins. A number of studies have found that among the various discretization methods, the method proposed by (Catlett, 1991) [Fayad et alia, 1991] using a minimum entropy heuristic is superior overall and is used in this project. Using r.reclass the new raster can then be generated whose categories are defined as ranges of the categories of the originating raster.

Predictors
The presence of a taxon at a selected locality is predicted using an ensemble predictors. Different predicting techniques are employed which include naive Bayesian and kriging methods, and

training is performed using different sets of features with the aim to produce accurate classifiers which can also disagree on their predictions.
Despite their simplicity, Bayesian networks are very successful learners, often better than more complex methods like neural nets and are becoming widely used (Heckerman, 1995).
Recent studies have also shown that the feature independence assumption can be relaxed somewhat. This had already become apparent in the success in practice of Bayesian methods even when, strict independence was not adhered to (Kuncheva et al., 2000) (Domingo, Pazzani, 1997) (Domingo,Pazzani, 1996)
The bayesian engine utilized is the GRASS program r.binfer.
The prior and conditional probabilities of taxa presence are assigned directly by local experts, inferred from ground truth and from the collocation of taxa on maps by the local experts and mined from Internet
Boosting (Freund, Schapire, 1999) or bagging (Breiman, 1996) can also be applied to each predictor to improve the performance.
Ground truth provides validation and refinement (Buntine, 1991a) (Buntine, 199b).
The different predictors are then assembled to produce a single predictor which should produce more accurate predictions than the single classifiers. (Opitz, Maclin, 1999) The integration of the single classifiers is implemented through weighted majority voting using a bayesian method (Shah and Aggarwal, 1999a) (Shah and Aggarwal, 1999b).

Performance Evaluation
The percentage of test examples correctly predicted is used as the estimate of predictor accuracy. When data is limited it is common practice to resample the data, that is, partition the data into training and test sets in different ways
Stratified random subsampling with paired t-test is used here to evaluate the accuracy of the predictors. In random subsampling (Geisser, 1975) (Schaffer, 1993) the data is randomly partitioned multiple times into disjoint training and test sets and the accuracies obtained from each partition are averaged. Stratification ensures that the class distribution from the whole dataset is preserved in the training and test sets. (Kohavi, 1995)
The present methodology of landscape mapping and biodiversity assessment will be validated through other independent surveys using traditional methodologies which will be performed in the area by visiting scientists.

PRELIMINARY RESULTS AND CURRENT STATES OF THE PROJECT
During the last year the premises to the present project have been validated.
The forest is very inhomogeneous at the scale utilized and the different habitats are apparent in the remotely sensed data utilized . The taxonomy and knowledge of biodiversity held by the local communities of the Rio Jauaperì are extensive . Local communities are willing to and capable of appropriating the new technologies.
Low end GPS equipment with amplified antenna has shown to provide sufficiently precise positioning even under thick tree cover, but only when at least four satellites are at least 45° over the horizon and well spread out. Image rectification with 5-10 meters accuracy seems possible and , in our case, sub-pixel accuracy consequently (Gao, 2001).
We have upgraded from GRASS 4.15 to GRASS 5.0 .

In October 2001 solar electricity, Internet connection and computers will be installed in the Xixuaù-Xiparinà Nature Reserve and a new impulse will be thus given to the project

The project is part of the International Biodiversity Observation Year 2001-2002 organized by DIVERSITAS.

REFERENCES

Gao, J. (2001) - *Non-differential GPS as an alternative source of planimetric control for rectifyng satellite imagery* - Photogrammetric Engineering and Remote Sensing - 67: 49-55

Minghelli-Roman, A; Mangiolini, M; Petit, M; Polidori, L. (2001) - *Spatial Resolution Improvement of MeRIS Images by Fusion with TM Images* - IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, 39:1533-1536

Fabbro, L. (2000) - *Amazonia Biodiversity Estimation using Remote Sensing and Indigenous Taxonomy* - ERS-Envisat Symposium "Looking down to Earth in the New Millennium" SP-461, European Space Agency Publication Division

Hall. A. M; Holmes, G. (2000) **-** *Benchmarking Attributes SelectionTechniques for Data Mining* **-** Department of Computer Sciences, University of Waikato, Hamilton, New Zealand

Kuncheva, L. I; Whitaker, C.J; Shipp, C.A; Duin, R.P.W. (2000) - *Is Independence Good For Combining Classifiers?* - Proceedings of the 15th International Conference on Pattern Recognition (ICPR2'00), 168-271

Qian Tan, Hui Lin, Yongchao Zhao, Tong Qingxi, Zhen Lanfeng (2000) - *Vegetation Spectral Feature Extraction Model* - Proceedings of The 21st Asian Conference on Remote Sensing, Hyperspectral and Data acquisition Systems, 1-7

Freund, Y; Schapire, R. E. (1999) - *A Short Introduction to Boosting* - Journal of Japanese Society for Artificial Intelligence, 14(5):771-780

Opitz, D; Maclin, D. (1999*)* - *Popular Ensemble Methods: An Empirical Study* - Journal of Artificial Intelligence Research , 11: 169-198

Shah, S; Aggarwal, J. K. (1999a) - *A Bayesian Framework for Multifeature/Multisensor Integration -- Automatic Target Detection and Recognition* - Annual Conference on Information Sciences and Systems

Shah, S; Aggarwal, J. K. (1999b) - *Statistical Decision Integration Using Fisher Criterion* - Proceedings of Second International Conference on Information Fusion, 722-729.

Zhukov, B; Oertel, D; Lanzl, F; Reinhäckel, G. (1999) - *Unmixing-Based Multisensor Multiresolution Image Fusion* - IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, 37:1212-1226

Domingos, P; Pazzani, M. (1997) *On the optimality of the simple Bayesian classifier under zero-one loss* - Machine Learning, 29:103-130

Williams, P.H; Humphries, C.J; Gaston, K.J. (1997) - *Mapping biodiversity value worldwide: combining higher-taxon richness from different groups* - Proceedings of the Royal Society, Biological Sciences, 264:1-24

Breiman L. (1996) - *Bagging Predictiors* - Machine Learning - 24(2): 123-140

Domingos, P; Pazzani, M. (1996) - *Beyond independence. Conditions for the optimality of the simple Bayesian classifier* - Machine Learning: Proceedings oftheThirteenth Internationak Conference of Machine Learning, Morgan Kaufman

SBSTTA (1996a) - *KNOWLEDGE, INNOVATIONS AND PRACTICES OF INDIGENOUS AND LOCAL COMMUNITIES*, - Second Meeting, Montreal, 2 to 6 September 1996, UNEP/CBD/SBSTTA/2/7

SBSTTA (1996b) - *ASSESSMENT OF BIOLOGICAL DIVERSITY AND METHODOLOGIES FOR FUTURE ASSESSMENTS* - Second Meeting, Montreal, 2 to 6 September 1996, UNEP/CBD/ SBSTTA/2/2

Heckerman D. (1995) - *A Tutorial on Learning Bayesian Networks* - Technical Report MSR-TR-95-06, Microsoft Research Advanced Technology Division, Microsoft Corporation

Kohavi, R. (1995) - *Wrappers for Performance Enhancement and Oblivious Decision Graphs* - PhD thesis, Stanford University

Williams, P. H; Gaston, K. J. (1994) - *Measuring more of biodiversity: can higher-taxon richness predict wholesale species richness ?* - Biological Conservation, 67: 211-217

Williams, P.H; Humphries, C.J; Gaston, K.J. (1994) - *Centres of seed-plant diversity: the family way* - Proceedings of the Royal Society, Biological Sciences, 256: 67-70

Fayyad, U. M; Irani K. B. (1993) *Multi-interval discretization of continuous-valued attributes for classification learning* . Proceedings of the Thirteenth International Joint Conference on Artificial Inteligence, Morgan Kaufman, 10022-1027

Schaffer, C. (1993) - *Selecting a classification method by cross-validation* - Machine Learning, 13:135-143

Buntine W, (1991a) - *Learning Classification Trees* - Artificial Intelligence Frontiers in Statistics, Chapman and Hall, London, 182-201

Buntine, W. (1991b) - *Theory Refinement on Bayesian Networks* - Proceedings 7th Conference on Unceratainty in Artificial Intelligence, Morgan Kaufman Publications Inc. 52-61

Catlett, J. (1991) - *On changing continuous attributes into ordered discrete attributes* - Proceedings of the European Working Session on Learning, Springer-Verlag, 164-278

Biederman, I. (1985) - *Human image understanding: Recent research and a theory*. Computer Vision, Graphics and Image Processing, 32:29-73

Geisser, S. (1975) - *The predictive sample reuse method with applications* - Journal of the American Statistical Association, 70(350):320-328