

GeoDMA

A novel system for spatial data mining

Thales Sehn Korting, Leila Maria Garcia Fonseca
Maria Isabel Sobral Escada, Felipe Castro da Silva
Image Processing Division, National Institute for Space Research – INPE
São José dos Campos, SP, Brazil
{tkorting,leila,isabel,felipe}@dpi.inpe.br

Marcelino Pereira dos Santos Silva
Department of Informatics, Rio Grande do Norte State University – UERN
Mossoró, RN, Brazil
marcelinopereira@uern.br

Abstract

Although a huge amount of remote sensing data has been provided by Earth observation satellites, few data manipulation techniques and information extraction in large data sets have been developed. In this context, the present paper aims to show a new system for spatial data mining, and two test cases applied to land use change in the Brazilian Amazon region. We present the operational environment named GeoDMA, developed to implement such approach.

1 Introduction

During the previous decade, significant progress has been made in planning and launching satellites with instruments suited for Earth observation. In addition, many remote sensing databases were built. While remote sensing provides quick and comparatively inexpensive information about land use over large areas, information extraction in remote sensing databases requires adequate methods. An example of a large database is the Brazilian Amazon deforestation data provided by PRODES project at INPE¹ [3]. Once the quick deforestation process causes land degradation, social tension, and irregular urbanization, faster and more precise identification of areas with these tendencies can reduce the consequences of such processes.

Data mining tools can, in fact, increase the analysis potential of such huge strategic data. However, few techniques for image data mining and information extraction in large

image data sets have been developed. Although there has been a large research effort in content-based image retrieval (CBIR) techniques, the specific problem of mining remote sensing image databases has received much less attention. Most proposals focus on clustering methods that operate on the feature space, which keeps the semantic gap between satellite images and smart extraction of strategic data.

Therefore, this paper aims to present a novel system for spatial data mining, called GeoDMA – Geographical Data Mining Analyst, which runs as an add-on for TerraView software, available at <http://www.dpi.inpe.br/terraview/>. TerraView is able to deal with spatial datasets, comprising images, and regions (or shapes) resultant from segmentation process. With such data (images and shapes), GeoDMA becomes able to extract several features, from spatial to statistical and spectral attributes, performing the complete data mining process, including attributes selection, training, classification, visualization and validation.

This paper is summarized as follows. Section 2 shows a brief review for data mining systems and approaches with spatial databases. Section 3 presents the GeoDMA system, showing its main aspects. Some results delivered by the system are shown on Section 4, followed by Section 5, where we conclude.

2 Spatial Data Mining in Remote Sensing Images

Recently, [13] has proposed a methodology to provide guidance for mining remote sensing image databases. Their method is applied to identify land use patterns in the Brazilian Amazon region from INPE's databases. To develop,

¹National Institute for Space Research – <http://www.inpe.br/>.



Figure 1. Spatial patterns of tropical deforestation (from left to right): corridor, diffuse, fishbone, and geometric [6].

test and validate the proposed methodology, a prototype called PattFinder (Pattern Finder) was built. PattFinder uses functionalities of different softwares such as SPRING [2], Fragstats [7], and WEKA [15] to implement each step of the entire process.

Other proposals, such as VISIMINE [1], ADaM [10], and KIM [11] are focused on clustering methods that operate on the feature space, created by the different spectral bands of a remote sensing image. These techniques are useful for distinguishing spectral signatures of different land use types, such as finding areas that are classified as “lakes”, “cities” or “forests”.

[13] proposed a methodology for mining large remote sensing databases using the idea of an application-dependent structural classifier. The methodology consists of three steps:

1. Definition of a spatial pattern typology according to the user’s application domain (Figure 1);
2. Building a reference set of spatial patterns. This reference set is built using a prototypical set of images. Landscape objects are identified and labeled: identification employs image segmentation and labeling is performed according to the spatial pattern typology (Figure 2);
3. Mining the database using a structural classifier (guided by the domain application), matching the reference set of spatial patterns to the landscape objects identified in images, thus revealing the spatial configurations present in each image.

The structural classifier distinguishes among different spatial patterns. This proposed methodology uses the C4.5 classifier [9], a classification method based on a decision tree. To select the attributes that distinguish the different types of land use patterns, the concepts from Landscape Ecology [14] were used. By applying the pattern metrics proposed by Fragstats² software [7], including Perimeter, Area, Perimeter-Area ratio, Shape index, Fractal dimension index, Related circumscribing circle, Contiguity index and

²Fragstats – Spatial Pattern Analysis Program for Categorical Maps

Radius of gyration, the system becomes able to characterize deforestation patterns in the Brazilian Amazon.

The landscape ecology metrics are fed into the C4.5 classification algorithm to distinguish the different types of spatial patterns. After the classifier is properly tuned, it can be used to label the landscape objects found in other images. Therefore, for each image in the database, this procedure identifies the number and location of the different types of spatial patterns. We refer to a specific set of spatial patterns found in an image as a spatial configuration. By identifying the spatial configurations of different images, the user is able to evaluate the emergence and evolution of different types of change.

3 GeoDMA System for Spatial Data Mining

GeoDMA system put together both approaches showed in previous Section. The input is composed by objects resultant from segmentation process (Figure 3), and also by the images used by segmentation. So we gather spectral and spatial information into the mining process, allowing the user a higher quantity of features for training and classification in order to classify and recognize patterns.

Figure 4 shows a diagram that explains the data mining steps included in GeoDMA system. One must consider that Image and Region Databases are provided by TerraView structure, and resultant Thematic Databases are also inserted into TerraView for visualization and storage for further analysis.

Now we describe each item presented on Figure 4:

- a) This module performs attribute extraction considering images and shapes by inputs. So, spectral features such as mean per band, covariance matrix and texture are extracted. And spatial features such as area, perimeter,

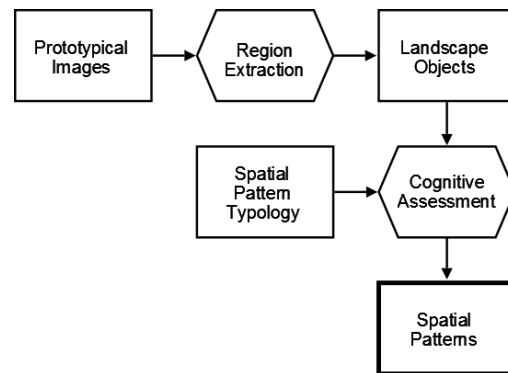


Figure 2. Building a reference set of spatial patterns.

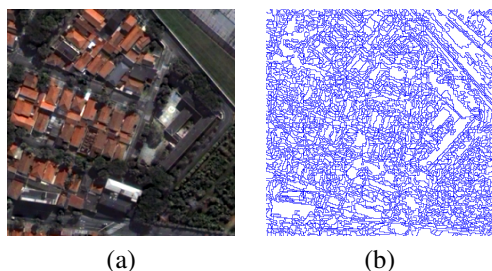


Figure 3. Image segmentation: a) Input and b) Output.

fractal dimension, compactness, rectangularity and main angle are also calculated and stored for further use;

- b) With this feature set, the user has an important tool for exploratory analysis, that is visualization in feature space, where two features are selected and one scatterplot shows the separability intrinsic to the data. Figure 5 shows one example of scatterplot using features `object_perimeter_per_area` and `pixels_mean_at_band_1`, in GeoDMA user interface;
- c) Normalization stage is an important step when performing exploratory analysis, since different scales can weight more one feature than other. GeoDMA provides two methods for normalization: linear $[0, 1]$ and standard deviation $[-1, 1]$;
- d) Two algorithms for supervised classification are available, and the source-code allows to develop new ones with a simplified interface. C4.5 algorithm and Self-Organizing Maps [5] are already implemented, and perform pattern classification considering input features selected by the user;
- e) TerraView user interface provides data visualization in a structure where the output can be splitted into different classes, considering classification results.

4 Results

This section shows one application in which GeoDMA was successfully used. It deals with land use dynamics [12], and have been applied to Amazon deforestation database provided by PRODES project. Extensive fieldwork also points out the different actors involved in land use change (small-scale farmers, large plantations, cattle ranchers), which can be distinguished by their different spatial patterns of land use [6]. The patterns evolve as time goes by; new small settlements emerge and large farms increase their agricultural area at the expense of the forest.

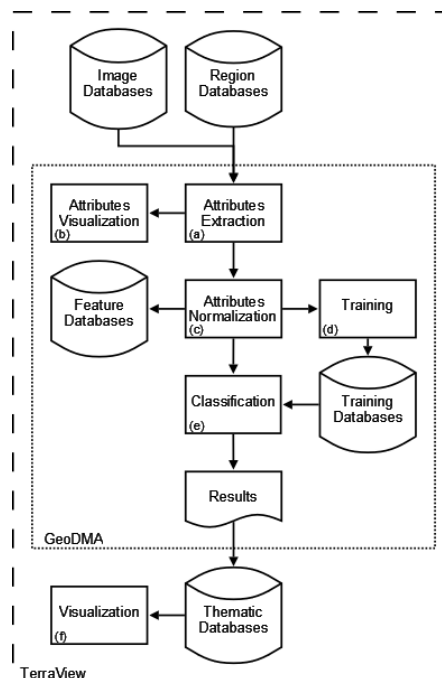


Figure 4. GeoDMA diagram – from images and regions to thematic maps.

[12] used GeoDMA to evaluate the landscape dynamic in Colniza municipality, based on deforestation pattern associated to different types of actors and their land use strategies. Thus, sizes and shapes of deforested areas are associated with different actors. Some classes of deforestation were created (such as Linear, Irregular, etc.) and they were associated to different kind of actors (Small Families, Rich Farmers and so on). Figure 6 shows some results. Figures 6a and 6b illustrate the deforestation patterns in 2005 and 2006.

Some advantage of using this system can be pointed out, as the facility to extract, normalize and store features for further use. In large datasets this feature becomes an useful

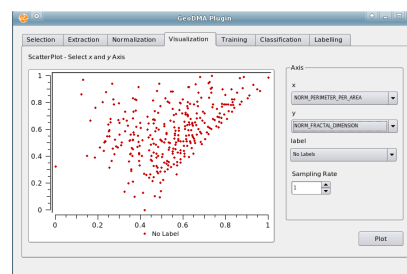


Figure 5. Scatterplot using spectral and spatial attributes.

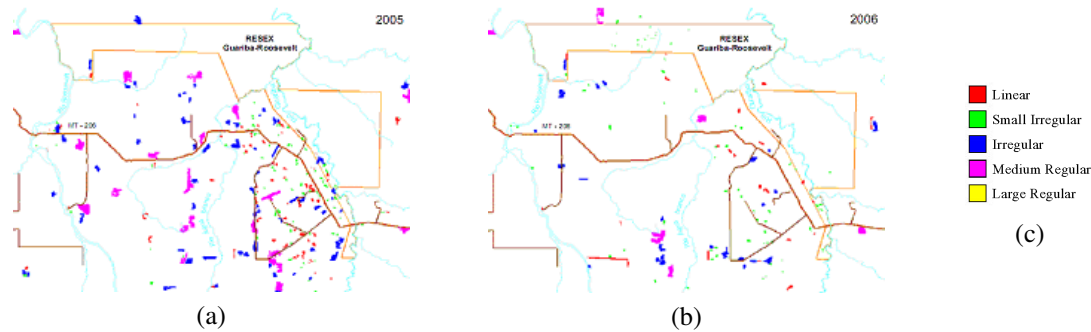


Figure 6. Deforestation patterns: a) 2005, b) 2006, and c) themes.

tool, specially considering workload and processing time.

Also the user interface provided by TerraView in which different patterns can be shown in separate, and the wizard-like interface of GeoDMA, where the data mining process can be performed without much expertise.

5 Conclusions

GeoDMA system was presented in this article. Similar methods for spatial data mining were analyzed, and functionalities from GeoDMA were exposed in detail. By exploring spatial and spectral features together, the system can reach better results, since the concept of neighborhood, intrinsic to the geographical analysis, is explored. GeoDMA also incorporates features for best attribute set selection, using metrics suggested in the literature [8, 4]. Using TerraView interface, GeoDMA becomes capable of storing every mining operation in the local database for further access, providing an easier way of working. The system can deal with all types of spatial applications where the input is described by images and regions.

Future works include optimizing the source-code, totally developed using TerraLib library³, specially the classification algorithms, when dealing with huge data volume. Another important functionality to be added to the system is the concept of multi-temporal analysis. The recovery of object evolution history aims to answer important questions about the causes of change. An understanding of historical land-use and land-cover patterns provides ways to evaluate pattern evolution thus helping to project future trends of human activities. More information about GeoDMA can be found at <http://www.dpi.inpe.br/geodma/>.

References

- [1] S. Aksoy, K. Koperski, C. Tusk, and G. Marchisio. Interactive training of advanced classifiers for mining remote sensing image archives. *Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 773–782, 2004.
- [2] G. Câmara, R. Souza, U. Freitas, and J. Garrido. Spring: Integrating remote sensing and gis by object-oriented data modelling. *Computers & Graphics*, 20(3):395–403, 1996.
- [3] G. Câmara, D. Valeriano, and J. Soares. Metodologia para o Cálculo da Taxa Anual de Desmatamento na Amazônia Legal. *INPE, São José dos Campos, Brazil. Available at http://www.obt.inpe.br/prodes/metodologia.pdf (retrieved 2005-08-10)*, 2004.
- [4] U. Fayyad and K. Irani. The Attribute Selection Problem in Decision Tree Generation. *AAAI-92: Proceedings Tenth National Conference on Artificial Intelligence/July 12-16, 1992*, 1001:48109, 1992.
- [5] T. Kohonen. *Self-Organizing Maps*. Springer, 2001.
- [6] E. Lambin, H. Geist, and E. Lepers. Dynamics of land-use and land-cover change in tropical regions. *Annual Review of Environment and Resources*, 28(1):205–241, 2003.
- [7] K. McGarigal and B. Marks. *FRAGSTATS: spatial pattern analysis program for quantifying landscape structure*. Portland, Or.: US Dept. of Agriculture, Forest Service, 1995.
- [8] J. Oliveira, L. Dutra, and C. Rennó. Aplicação de Métodos de Extração e Seleção de Atributos para Classificação de Regiões. *XII SBSR*, pages 4201–4208, 2005.
- [9] J. Quinlan. *C4. 5: Programs for Machine Learning*. Morgan Kaufmann, 1993.
- [10] J. Rushing, R. Ramachandran, U. Nair, S. Graves, R. Welch, and H. Lin. ADaM: a data mining toolkit for scientists and engineers. *Computers and Geosciences*, 31(5), 2005.
- [11] M. Schroder, H. Rehrauer, K. Seidel, and M. Datcu. Interactive learning and probabilistic retrieval in remote sensing image archives. *Geoscience and Remote Sensing, IEEE Transactions on*, 38(5 Part 1):2288–2298, 2000.
- [12] F. C. Silva, T. S. Korting, L. M. G. Fonseca, and M. I. S. Escada. Deforestation pattern characterization in the Brazilian Amazonia. *SBSR*, 2007.
- [13] M. Silva, G. Câmara, R. Souza, D. Valeriano, and M. Escada. Mining Patterns of Change in Remote Sensing Image Databases. *Proceedings of the Fifth IEEE International Conference on Data Mining*, pages 362–369, 2005.
- [14] M. Turner. Landscape Ecology: The Effect of Pattern on Process. *Annual Reviews in Ecology and Systematics*, 20(1):171–197, 1989.
- [15] I. Witten and E. Frank. Data mining: practical machine learning tools and techniques with Java implementations. *ACM SIGMOD Record*, 31(1):76–77, 2002.

³Available at <http://www.terralib.org/>.