

**GEOPROCESSING IN DIAGNOSTIC AND PROGNOSTIC OF AREAS OF SPECIAL INTEREST
IN THE AREA OF INFLUENCE TRANSMISSION LINES OF CEMIG - A CASE STUDY OF
METROPOLITAN AREA IN BELO HORIZONTE - MG - BRASIL**

MOURÃO A.(1), SANTANA S.(2), MOURA C.(3), LANNA L.(3), AZEVEDO U.(3), LOURENÇO P.(3)

(1) Federal University of Minas Gerais, BELO HORIZONTE, BRAZIL ; (2) COFFEY Information, BELO HORIZONTE, BRAZIL ; (3) CEMIG, BELO HORIZONTE, BRAZIL

ABSTRACT

The study aims at the exploration and optimization of geospatial technologies in the characterization and analysis of the use and occupation of land near transmission lines Cemig's electricity, with a view to identifying areas of priority interest for monitoring the dynamics of occupation and as for the construction of predictive studies of the possibilities of transformation in these areas. Part of the project actions Geomap, which includes other actions to build detailed databases about the area of transmission lines, the actions reported in this article refer to planning actions on the region of transmission lines. It promotes broad-based structuring of data, with variables that characterize environmental conditions and the anthropogenic land cover in the areas of transmission lines in the metropolitan region of Belo Horizonte, among which we mention the infrastructure, density and type of occupation, topographical conditions, accessibility factors, socio-economic areas of urban expansion, among others. It structures Geographic Information System for queries and spatial analysis. From the collection of data the models identify areas of occurrence of space activities, studies of correlations of variables and combining variables. Apply models of spatial analysis based on Multicriteria methodology for addressing procedures guided by experts (knowledge driven evaluation) and statistical procedures according to the behavior of the data (data driven evaluation), including the Expert System, the Signature and Data Mining (Data Mining). By relying on the study of methodological exploration of multi-criteria analysis aiming to expand its application to the entire state of Minas Gerais, generates four summaries in four different methodological procedures, in order to validate results and identify the procedures less investment in time and operating cost. The results are distribution maps of potential areas of interest to the comments and attention from the company, as well as the characterization of conditions for the use.

KEYWORDS

Geoprocessing, Multicriteria analysis, diagnostic and prognostic studies, LTs

INTRODUCTION

The study aims at the exploration and optimization geospatial technologies in the characterization and analysis of the use and occupation of land near transmission lines for electric power CEMIG with a view to identifying areas of priority interest for monitoring the dynamics of occupation, so as to build predictive studies of possibilities of transformation in these areas. The project, called Geomap, encompasses other actions to build databases detailed region of the transmission lines, with the support of geo-edge for development of high-precision registers. As part the larger project, the actions reported by this article refer to planning actions on the region of the transmission lines.

The project aims to Geomap report generation and characterization of the areas of transmission lines for planning and management. It is understood as a set of management actions in smaller regional and temporal: it is the daily, based on entries and detailed intervention procedures. On the other hand, the planning actions aimed at larger territorial and time for setting goals and goals. It is understood that both actions, planning and management, are key to company, and the first defines the major goals and objectives, and establishes a second act as locally and to get immediate results expected. The action planning approach, especially the study of the conditions of the tracks field of transmission lines and their surroundings immediate, more specifically 250 meters distance to each side of the shaft. Procedures include identification of the invaded areas and areas susceptible to invasion by others, defining areas of attention. The characterization of the areas was obtained the mapping of variables, with the support of GIS on the issues of land use, topography, density of occupation, road system, topographic conditions, vegetation conditions socio-economic, among others. The methodology was based on building cartographic databases and information integration through the algebra of maps known as "Analysis of Multiple." As the study area, was initially selected a pilot area within Belo Horizonte, which was built for the script methodology. Once performed the first study, project was extended to the metropolitan area, with necessary adjustments to increase the spatial complexity of the area.

BASIC CONCEPT - ANALYSIS MULTICRITERIA

According to Moura (2003) Analysis Advanced is a methodological procedure of variables, including the widely accepted spatial analysis. She is also known as Hierarchical Analysis of Weights. The procedure is based on the mapping of variables per plan information and defining the degree of relevance each plan for information and each of its legend components for building the result final. The mathematics employed is the simple average Weighted. The use of Weighted Average creates a space classification, ordinal, which can also be understood as an interval scale. This process can also be used in nominal scale, since that events are ranked according to some value criterion. The balance must be made by knowledgeable of the phenomena and the variables situation assessed, or by prior knowledge of similar situations. In this process, the possibility to improperly consider a situation is inverse of the number of weightings. Delphi (or the query specialists) in obtaining the weights and grades based on in choosing a multidisciplinary group of specialists, who know well the phenomenon better and even if they know well the reality of space where he is located. To these experts are asked to please rank or place variables (or plans information) in order of importance for event or occurrence of phenomenon. Another way of assigning weights is building based on statistical analysis of to identify situations in which there is low correlation between variables and high correlation with the phenomenon or occurrence to be explained. This is because if there are high correlations between variables, both would be contributing equally to the explanation of the phenomenon (is like a substitute to another analysis, and incorporation of both would not difference). However, it is important to note that although there are methods and techniques to take the specialist assign responsibility for all numerical values analysis, represented by the weights and grades of information plans and their respective legend components, there will always be a level of subjectivity. There will always be a need to indicate a hierarchy between variables (which are added in 100%) or the weights of its subdivisions (receiving membership degree of 0 to 10). There is no research a small trace of subjectivity, because the very choice of a model already is the opinion of an expert a second look at the spatial reality.

MAPPING SPECIAL FEATURES BANDS IN AREA POWER LINES TRANSMISSION ELECTRICITY IN METROPOLITAN AREA OF BELO HORIZONTE - MG - BRASIL

The assembly of the database follows the following script:

- a) Definition of variables mapping sources and available data and new data to be generated;
- b) Definition of mapping scales each variable;
- c) the choice of territorial unit data integration, which means the size of pixel matrix definition and representation.
- d) Resampling of information layers resolution set to in the preceding item;
- d) Assembling the cartographic database and alphanumeric

Once made up the database, which are information plans in the form of maps, tables and images, is structured GIS (Information System Geographic), according to the script:

- a) Association of alphanumeric data and composition of thematic consultations of interest to study and process modeling;
- b) Assembly containing the GIS plans information in different formats for archiving (Vector or raster), but especially with plans that will be used in models of spatial analysis in raster (matrix);
- c) Construction of thematic maps for consultations and distribution of phenomena in the territory.

Since structured GIS, is applied spatial analysis model called "Analysis Advanced "according to the script:

- a) Verification of the choice of variables mapping - from experts and conference cognizant of the phenomenon on the possibility of insertion of new information plans;
- b) Application of Delphi method to ranking of the weights of the analysis variables;
- c) Application of statistical method Hierarchical Analysis of Weights, Expert System (Specialist) and Data Mining (Data Mining) for assigning weights to variables.
- d) Verification and validation of results obtained by checking the data check the high resolution image or compared to familiar situations.
- e) The products generated were characterized zonings under different environmental variables, with identification of special situations characterize the study area, the second conflict, potential risks and priorities for intervention.

Composition Database

We worked databases vector alphanumeric (tables) and high resolution images overlooking the junction of the variables and construction maps of potential interest in space. The first step of every project in the area of GIS is to gather the basic data available and match them. As we are working with data produced by

various institutions, was necessary to standardize the units of integration, map projection and scale of data worked. The projection was chosen as the UTM, Sad 69, Zone 23. Some details were worked by census, data points, linear data (Infrastructure), pixel (picture) and others from digital elevation model resulting from the work of Extraction of Digital Terrain Model from ASTER satellite images. Layers of information generated:

a) Distribution Infrastructure: Service Water Supply Network for General Service Sewerage Network for General Service Garbage Collection Service for Public Service Electricity Supply. Information were represented by census tract or surface distribution of services (maps Kernel) identifying the areas best served by the most poorly served by infrastructure in the region.

b) Accessibility Road: map built From the classification of the concentration of urban roads and roads, the latter being separated into paved and unpaved. Information were constructed from images vectorization satellite high-resolution, QuickBird. From highways have been mapped buffers, which are bands of distance, so that the surface generated allows assess the degree of accessibility at each point territory.

c) Urban Services: Distribution trade, services, industry and services collective use from data associated with the poles of CEMIG. The Kernel density was used to generate proximity to the surface distribution of one or most urban services. To sort the branches of activities within these four groups was used as a reference table provided by the Plan Director of Law and Installment, Occupation and Use of Solo de Belo Horizonte.

d) Population density and Deployment Buildings: population distribution and buildings in the area. The data of population density was worked from the census and table from the 2000 Census. Since the density of construction of the buildings has been obtained from the classification of high-resolution images.

e) Slope (Topographical Conditions):The slope was extracted from satellite images ASTER. The procedure promoted the construction of digital elevation model and further development of thematic map in the slope levels out to the characterization of topographic conditions on tracks that follow the Federal Law 6766/79, which defines areas restriction to urban occupation.

f) Socioeconomic were mapped the spatial distribution of average income family and schooling, from both sectors and census data from IBGE.

g) Proximity to Villages and Slums: variable mapped by classification of images of high resolution and the construction of buffer strips or domain of these occurrences.

h) Pattern of Vegetation: Mapping was done through the separation of vegetation types trees, shrubs, scrub, cultivation and soil exposed to from the classification of the QuickBird images.

i) Use and Land Cover: mapped through the vectoring of high resolution images (QuickBird, Orthophotos and WordView), according to sort keys: slums, high density urban average urban density, low density urban major landscape structures, exposed soil, undergrowth, trees, vegetation shrubs, cultivation, mining substations Cemig watercourses and roads.

j) Area of Urban Expansion: obtained through Classification and comparison of satellite images CBERS 2001 and 2007.

k) Proximity to the city limits: constructed from the classification of images Satellite CBERS 2007. It is important to remember that concept of the urban area is the division politic administrativa arbitrated by the Municipal Councils of so that the work was aimed at identifying contacts concentration of anthropogenic occupations.

l) Proximity to substations: classification use and land cover in the QuickBird images.

m) Proximity to Large Structures Landscape: vectorized from the images of high resolution. They are composed of large warehouses, parking lots, football fields, clubs, industries, schools, hospitals and other structures that stand out the landscape for its power of attraction.

n) Surveillance: Maps of routes inspection of transmission lines and their periodicities.

Multicriteria Analysis

The multicriteria analysis is the combination of variables for process algebra maps, aiming to the composition of a synthesis that translate some potential that one wishes to investigate.

Application of the Delphi method to integrate variables - Knowledge Driven Evaluation

The Delphi method in obtaining the weights and notes based on the choice of a group multidisciplinary experts, who know well the phenomenon and even better to know each other well spatial reality where it is located. To these experts are asked to place in rank or put variables (information or plans) in order of importance for the manifestation or occurrence of the studied phenomenon. Example: for generation map of potential risks, what is the order of importance of chosen? Upon receipt of responses from the group,

carried out the selection average and an indication of the predominance manifestations. The specialist then receives the result consultation and asked to review their positions – if He has the firmness of your choices, keep their answers, but if he decides to adjust its ratings before the group's response, he expressed new opinion. So it is done for two rounds, but there are situations they apply three rounds. (Moura, 2009). It was initially held lecture explaining the method, its procedures and objectives, illustrated with the example of pilot developed in the previous project. From this lecture CEMIG structured the list of people who would participate in the Delphi questionnaire, which aims to maximizing consensus and consider different views on the subject. These participants received a e-mail stating the relationship of variables covered in the pilot, from which they should asked about the maintenance of those variables, as well as the inclusion of new variables. Organized a list of variables from the justification for their uses and checks conditions for the realization of this mapping. The List, the second step was to ask participants thinking about the importance of each variable for the composition of an area of interest special observation by the company, indicating values from 0 to 10, with 0 indicating that the variable would have minimal significance and 10 that have high variable observed. Participants were given the table resulting from step 2, without identification of the other participants, and containing the mean and mode of opinions. The third stage of the process was to check whether the views they held or would like reviewing suggestions from the information mean and median. About half of the participants adjusted their opinions. The goal is to avoid being Grading at random and that person has opportunity of knowing the group position themselves in relation to the majority. Once received the opinions of review the participant, new media and fashion are computed and is drawn up the final table of values assigned to each variable. Held this synthesis is applied slicing of the values for the statistic of "Break- Natural "(Natural Breaks) so that the set of variables is divided between those that have high medium to high, medium, medium low or low importance, according to experts, about the interest to special care by the company The variables indicated by the group were structured in 20 layers of information, and transformed into data arrays configured as surface potential distribution of each variable. According to Moura (2009) the first step is to composition of a cartographic database, composed in the form of plans and information that should be combined with the application of models spatial analysis can be done in vector format or matrix, but there are strong trends towards predominance of the operations of the models in formats matrix (raster). The question is justified by the relationship topology implicit in the process matrix, which is not only optimizes the exchange of data, but also sine qua non in some models.

It is important to define the territorial unit integration of the analysis, which means the choice of resolution of information layers, and consequently, the spatial resolution or precision analysis generated. In this study the resolution space was 5 meters. The values of each position of the matrix, corresponding to a position territorial, were reclassified according to the degree of importance. Once built surfaces potential or layers of information they are synthesized by the weighted average process map algebra, as the following:

Where:

$$A_{ij} = \sum_{k=1}^n (P_k \times N_k)$$

Aij - a position in the array analysis (Row / column), or pixel map;

n - number of maps or layers of variables cross;

Pk - percentage points or weight assigned to map or layer of variable k;

Nk - degrees of influence (from 0 to 10) of type of variable to the final risk assessed

The use of Weighted Average creates a space classification, ordinal, which can also be understood as an interval scale. Was structured table that lists the variables combined and their respective weights, and the product of crossing was a general map of area classification, shown in Figure 1

Application of the Data Mining and Signature - Data Driven Evaluation

The method of combination of variables for Advanced indicate two ways of obtaining weights and notes to indicate the degree of relevance of each one of them throughout the synthesis: the Knowledge Driven Evaluation (by expert opinion) and Driven by Data Evaluation (by the behavior of data). The processes are not mutually exclusive, but complementary, since they result in different views on the same study.

Knowledge Driven Evaluation in order is drawn from the experience of experts indicating the list of variables of interest for the phenomenon, so the behavior of the variables. There are different methods to obtain the synthesis of the views of experts with the objective of maximizing search consensus. In the

present study was chosen Delphi method for the synthesis of consultation and Data from interviews with experts. In Data Driven Evaluation is the goal extract data from a list of variables that make areas of interest for observation due to risk of invasion. In our study, this process was conducted on two approaches: Signature in squatter areas and an approach for Data Mining, for the validation of final results. Signature:

We identified 254 locations invasions confirmed, and there were marked points these occurrences. Likewise, we 254 selected other locations where it was confirmed the non-invasion, and marked these points occurrences. Once built the collection of points, were built buffers (areas of influence) of 15 meter radius or 30 feet across from them, and carried the signature on each layer to identification of the behavior of those variables locations. The result is the dominance by area of each component caption of the variables mapped. The interpretation of the signatures on the table identifies which variables are most relevant for the occurrence of the phenomenon, as well as what legend components (characterization of the variables) are more relevant to the phenomenon. The interpreting the results of a signature must following procedures:

- Subscribe points or places where the phenomenon that interests us has been confirmed but is important to also sign the entire study area for know the behavior of variables in a way general.

Analyze the results of phenomenon always confirmed ahead of the general characteristics of the set.

- To establish a hierarchy of the importance of the variable or map to characterize the phenomenon. Observe that the phenomenon happens in any situation (if there are values similar in all the legends of the map) that variable, or if there is a predominance of one or more captions. When the distribution is homogeneous, that variable or map is not useful to highlight areas potential for the phenomenon, but is dominated legend components in the distribution, that variable or map should be considered.

Once selected variables or maps to be employed, the values in the signature should help us to assign weights and grades. It is made order of importance observing these values. It built new relationship table of weights for variables and performed a new synthesis, followed by spatial representation in the map. There is a significant similarity between the maps produced in first step: the evaluation of knowledge-driven (which followed the values indicated by Delphi) and the date driven evaluation (which followed the values indicated by signature, which reflects the behavior of the data in territorial reality studied). The similarity expressed that looks very experts are calibrated reality, being then further investigation details of where the differences occur in rankings and holdings of the variables (Figure 1)

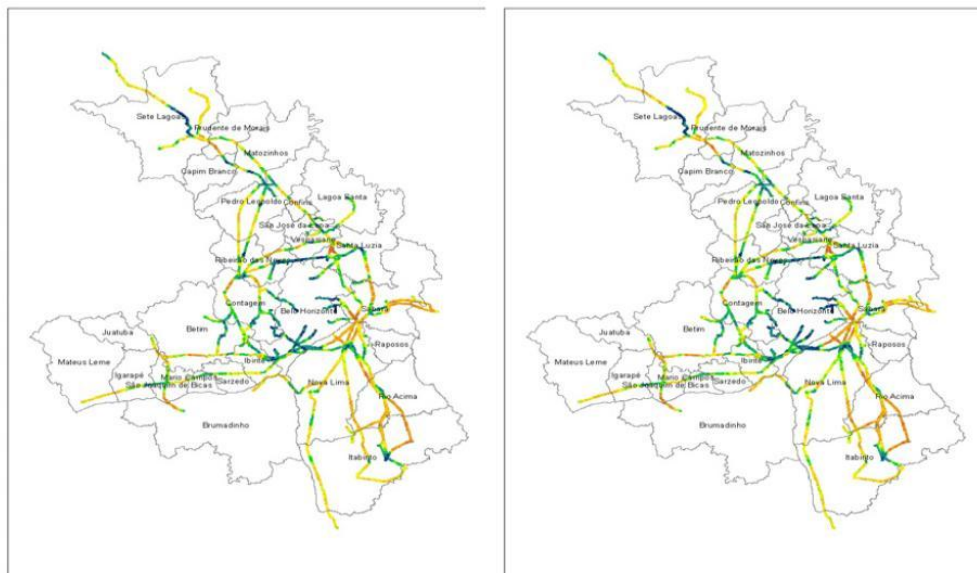


Fig 1: Left map obtained through the Delphi process and to the right of the map obtained by Subscription Data Mining

Since the goal of establishing the process decision in selecting the most important variables intervene more actively in the occurrence of the phenomenon studied, we propose to employ the logic of heuristic for data mining.

According to Xavier-da-Silva, communication oral October 24, 2007: "A heuristic is a way of looking at qualitative view of the combination variables that account for a certain phenomenon, since it allows different professionals opine about an environmental situation and adjust their views from the

understanding of the degree relevance of each component of environmental According to the context where it is inserted, so as the degree of relevance of his remarks ahead the thought of the group of other experts. " Among the heuristic procedures, adopt data mining to support the selection variables, in order to identify patterns relationship between them to justify the identification of associations that best answer for the phenomenon. Data Mining, in particular, is the step where they are algorithms applied toward achieving goals specific, producing an enumeration of particular patterns in the data (Goebel and Gruenwald, 1999). So in a complete process of discovery knowledge can be used several Data Mining algorithms. Data mining is a spatial extension of data mining oriented domains application where the consideration of the spatial dimension is the extraction of essential knowledge. So Similarly, there are developments of work considering the temporal context (Temporal Data Mining) and, considering both dimensions (Spatio-Temporal Data Mining). It is organized table that separates records squatter areas and records of areas not invaded, together with their characterization according to variable mapping. The table is treated with free software Weka (Waikato Environment for Knowledge Analysis), "Environment Waikato to Analysis of Knowledge ", developed by University of Waikato, New Zealand. He was selected the Bayes classifier (Naive Bayes) and the action of cross-validation (crossvalidation). Naive Bayes classifier suggests that presence (or absence) of a particular characteristic a class is not related to the presence (or absence) of any other characteristic. The classifier that considers all properties independently contribute to the probability of an occurrence (Zhang, 2001). The result shows the confusion matrix:

Table 1 - Confusion Matrix

	a	b
a	5423	1943
b	566	6800

The interpretation of the matrix indicates that there were inserted data type "a" - invaded area and data "b" - area not invaded. The whole "to" 5423 records were identified as having the predominant features of "a" and 566 were identified as having the characteristics predominant "b". Data set type "B" 6800 were classified as having predominant features of "b" and 1943 as with predominant features of "a". That is: system performs a classification and identifying patterns behavior, and then it checks the number of occurrences which he expected answer, by similarity of pattern, with the type "A" or "b". In the case in particular, we consider confusion of 10% in the classification of "overrun" is a very low error rate. The confusion became greater with the "no invasion" was greater, about 29%, which means that areas not conducive to the invasion have a greater complexity in its conformation. Was a the great result of the invaded areas, because it signals that there is indeed a set of similar variables that form a pattern that favors the occurrence. Another index that supports this observation is Kappa index that the system generates very used to give an idea of how much the observations deviate from those expected, children of chance, thus indicating how legitimate interpretations are (Jensen, 2005). Kappa index is calculated for each confusion matrix.

In the study gave the Kappa index 0.66, which places, according to the above table, the category of "Substantial", second only to the excellent. From the study of relationships between variables and pairwise by observing the standard deviation obtained for each new relationship was proposed for variable selection and their values, to perform new crossing Advanced. The aim of this step was to identify the main components, ie, the main variables that account for the phenomenon and once the cut group, to assess the degree of assertiveness that can get working with just those variables. It was proposed that the new crossing, with the reduction of number of variables.

CONCLUSIONS

The resulting map by data mining (data driven evaluation) was compared to the map result and the result was the signature validation results and indicated that it would be possible yes reduce the number of variables to the principal without great loss of information. Thus, in projects future can be reduced to 10 layers variables and arrive at results very similar as shown in Figure 2

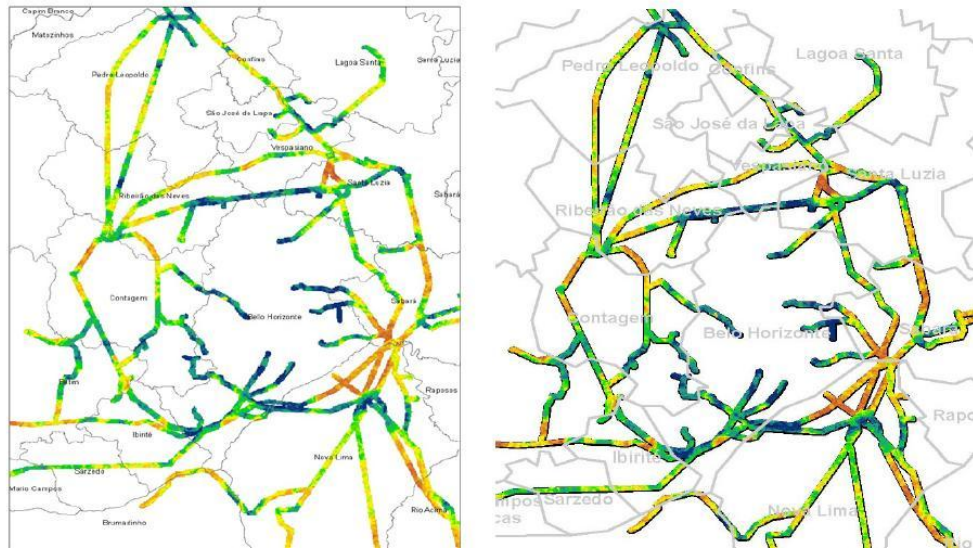


Fig 2: Left section of the map obtained by Signature and on the right obtained by data mining

Importantly, the resulting map Data Mining serves as validation of results towards reality, as they are applied indexes spatial correlation and the degree of recognition of behavioral patterns for the phenomenon from its main variables. The procedure is also a way to adjust the data to arrive at an index legitimacy of the procedures. Given the above, we consider the work satisfactory in the sense that included different methodological procedures for comparing results, and validation of the results presented obtained. From the validation of the methodology was promoted the intersection between the map overview final which ranks the country according to the degree of interest to special care and observation in the surveillance, and the map of land cover. The goal was to perform measurements on the event types that located in areas identified as high a average interest such as dense urban occupation, undergrowth vegetation, exposed soil, slums, among others. The results were validated by different methodological procedures and confirm the potentiality of the methodology and the techniques employed to support decision-making in planning areas of influence of power transmission lines

Electric CEMIG. The collection of data on RMBH municipalities is a significant collection that can and should be used in other projects company, since it addresses characterizations environmental conditions and human activities, infrastructure and conditions of land use.

BIBLIOGRAPHIC REFERENCES

- Bonham-Carter, G. Geographic information systems for geoscientists: modeling with GIS. New York: Pergamon, 1994.
- Goebel, M., Gruenwald, L. A survey of data mining knowledge and discovery software tools. SIGKDD Explorations, p. 20-33, 1999. In.: Neves, Marcos Correa Freitas, Corina Costa and House, Gilbert. Data Mining in Large Databases Geographical. INPE, Technical Report, November, 2001.
- Moura, Ana Clara M. Methodological Discussion applying the model on Voronoi Polygons studies of phenomena in the areas of influence urban occupations - a case study in Ouro Preto - MG. Anais VII ENAP, São Paulo, Brazil, 9-11 September 2009, FEA / USP.
- Moura, Ana Clara M. GIS in the Management and Urban Planning. Belo Horizonte, Ed's author, 2003.
- Xavier-da-Silva, J. GIS for analysis environment. New York: J. Xavier da Silva, 2001. 227 p.