Computer Technology and Application 5 (2014) 25-32

DAVID PUBLISHING

Spatial Data Mining to Support Environmental Management and Decision Making—A Case Study in Brazil

Carlos Roberto Valêncio, Fernando Tochio Ichiba, Guilherme Priólli Daniel, Rogéria Cristiane Gratão de Souza, Leandro Alves Neves and Angelo Cesar Colombini

Department of Computer Science and Statistics, São Paulo State University, São Paulo 15054-000, Brazil

Abstract: The growth of geo-technologies and the development of methods for spatial data collection have resulted in large spatial data repositories that require techniques for spatial information extraction, in order to transform raw data into useful previously unknown information. However, due to the high complexity of spatial data mining, the need for spatial relationship comprehension and its characteristics, efforts have been directed towards improving algorithms in order to provide an increase of performance and quality of results. Likewise, several issues have been addressed to spatial data mining, including environmental management, which is the focus of this paper. The main original contribution of this work is the demonstration of spatial data mining using a novel algorithm with a multi-relational approach that was applied to a database related to water resource from a certain region of São Paulo State, Brazil, and the discussion about obtained results. Some characteristics involving the location of water resources and the profile of who is administering the water exploration were discovered and discussed.

Key words: Water resource management, spatial data mining, multi-relational spatial data mining, spatial clustering, environmental management.

1. Introduction

Geographic data have been collected with the use of modern data collection techniques, like GPS (Global Positioning System), remote sensing and others [1], and the development of GIS (Geographic Information System) has resulted in new approaches to store, transmit and visualise modelling [2]. The development of geo-technologies has motivated the application of GIS and remote sensing in ecological areas in order to aid ecological system comprehension [3-8]. Evolution of these technologies has resulted in large scale geographic data repositories and this situation motivates the development of methods for information extraction. In this context, spatial data mining has emerged as a research area to address the challenges of this complex task [9].

Spatial data mining aims to identify useful patterns in large spatial databases using algorithms that implement some 48 methods like clustering, classification, association rules, spatial trends and others in order to explicit spatial information [10, 11]. This process is more complex than conventional data mining due to the complexity of spatial data, spatial relationships and spatial autocorrelation differently of traditional numeric and categorical data [12].

Several methods for spatial data mining and a formal categorisation of them have been proposed. Clustering is one of the most used approaches and the case study described in this paper was based on an algorithm of this category, which consists on identifying groups of spatial objects that are as similar as possible to each

Corresponding author: Carlos Roberto Valêncio, professor, research fields: database, KDD, spatial data mining. E-mail: valencio@ibilce.unesp.br.

other and, in comparison to those in another group, are as dissimilar as possible [10, 13]. Algorithms for clustering in spatial data mining can be classified into four subcategories: segmentation approach, hierarchical approach, density-based or grid approach [11].

In another category there are methods for spatial association rule mining, similar to association rule in relational database. Algorithms of this category seek for a rule that describes the situation in which a set of features implicate another set of features in spatial databases. The condition is based on the form A → B[s%, c%]—A and B are sets of spatial or non-spatial predicates; s% is the support of the rule; c% is the confidence of the rule [14].

Spatial classification refers to methods which group objects into classes, also named categories, by analysing their properties—it includes their spatial and non-spatial attributes, the spatial relationship to their neighbours and the attributes of their neighbours. In this process, the classification model needs to be trained using a training dataset to validate the configuration and evaluate the performance [15].

Considering the application of spatial data mining, several areas have been benefited in cases in which traditional data mining is not enough to provide useful information. One example of the use of spatial data mining techniques is the application in marine geographic information system, that supports the elaboration of a visualisation technology to represent true three-dimensional submarine topography, the analysis of changing rule of marine sediment according to the spatio-temporal element in order to provide safe ship navigation, the analysis of navigation data in order to optimise the program that aids navigation and the rule of hydrological elements to support navigational decision making [16]. Another case is a work that performs spatial data mining using literacy rates and educational establishments in Bangladesh in order to analyse the behaviour of variables according to spatial location and to compare the results to those obtained with a classical regression model. The authors used spatial autocorrelation, Moran Scatter plot and spatial regression and concluded that there is a spatial consistency in the distribution of literacy rates and educational establishments in Bangladesh, based on the results of the work [17].

Results of spatial data mining are interesting for several areas of application, but this research area has to overcome some challenges and limits: the computational cost of spatial data mining algorithm execution is high, so in some cases the application of these technologies is impracticable; the complexity of algorithms is high and sometimes is exponential; some approaches for spatial data mining are inherited from traditional algorithms for data mining, which present low quality of results and difficulty with the extraction of useful information. Thus, the efforts of researchers in this area are focused on the development of new approaches based on the context of spatial data, which is different to conventional data, in order to improve the quality of results and the required execution time [11, 18].

In this work, the algorithm used an innovative approach of density-based clustering that implements an improvement on the approach of DBSCAN (Density Based Spatial Clustering of Applications with Noise) [19] and VDBSCAN (Varied Density Based Spatial Clustering of Applications with Noise) [20] and increments it to the approach proposed by CLARANS (Clustering Large Applications based on Randomized Search) [21], a combination of techniques which presented better results in comparison with the approaches in an isolated way due to the consideration of spatial and non-spatial similarity in spatial clustering—a new method of data analysing. Besides this, it presents a novel approach for multi-relational spatial data mining, which guarantees semantic increasing of results and better performance in comparison to conventional approaches, once it avoids data joining before the data processing [22]. This technique was applied to a database related to natural resources from a region of São Paulo State, Brazil, in order to demonstrate the results and the kind of information it is able to provide for better decision making, once spatial data mining aims to survey relationships and characteristics—spatial or non-spatial—that may be present in a spatial database [23]. Results of this work demonstrated that information can be related to other well-known elements and also used to support managers for better decision making.

2. Material and Methods

Spatial data mining in the context of water resource management was performed using an algorithm which presents some specialties for execution of analysis in a multi-relational approach and implements improvements for quality, increasing results and performance, in comparison with conventional approaches [22]. The focused database is the result of a work that has been developed since 2010 in a region of northwest of São Paulo State, in Brazil, seeking to establish a survey about the situation of water resource exploration. The target area is considered critical, since the exploration of natural resources is over the limits of sustainable use. The major element that contributes to this situation is large scale urbanisation, which is nearing 1.5 million inhabitants—in a region of 66 municipalities [24]. One of the municipalities in the described region is called Votuporanga, where more than 83 thousand people live and where a fieldwork was executed to collect data related to water resources found there, visiting each one and collecting a large scale of data related to characterisation of water usage, method of exploration, water availability and others, besides the elements that describe the farms and their owners, like use of soil, type of erosion control, kind of predominant culture, destiny of rubbish, type of irrigation, scholarity level of those who administrate water exploration and others. It is important to highlight that visited water resources were geo-referenced, which enables the application of spatial data mining algorithms and the visualisation of their spatial distribution.

The fieldwork team filled out a form with collected data into a GIS-based system to support the environmental management [25] and the resulting database was used to execute the spatial data mining algorithm in this case study. In this repository, there are data related to more than 500 geo-referenced water resources that enable analyses considering the spatial and non-spatial features.

The distribution of water resources is shown in Fig. 1, in which each coloured little circle around the municipality of Votuporanga represents a geo-referenced water resource visited by fieldwork and that has its data set in the analysed database. This figure was taken from GIS-based web system and each circle has a colour which represents the destiny of the explored water.

In order to understand the representation in Figs. 2, 3 and 4, some information has to be enunciated: each red square corresponds to a certain water resource catchment location from database; polygon filled in green colour is the region that delimitates the cluster; and blue squares are water resources that were considered as part of the correspondent cluster.

3. Results and Discussion

The first application of spatial data mining methods resulted in a cluster presented in Fig. 2, in which 48 water resource catchment locations were grouped since they have some similar characteristics identified by the algorithm, in this case, all of them are administrated by people with a poor level of education. Considering that the entire database has 507 geo-referenced water resource catchment locations, the cluster in Fig. 2 represents 9.4% of all registers. Although this percentage is not representative, analysing this cluster in comparison to all water resource catchment locations that present the identified characteristics, in this cluster 73.8% of them are present—a more representative information.

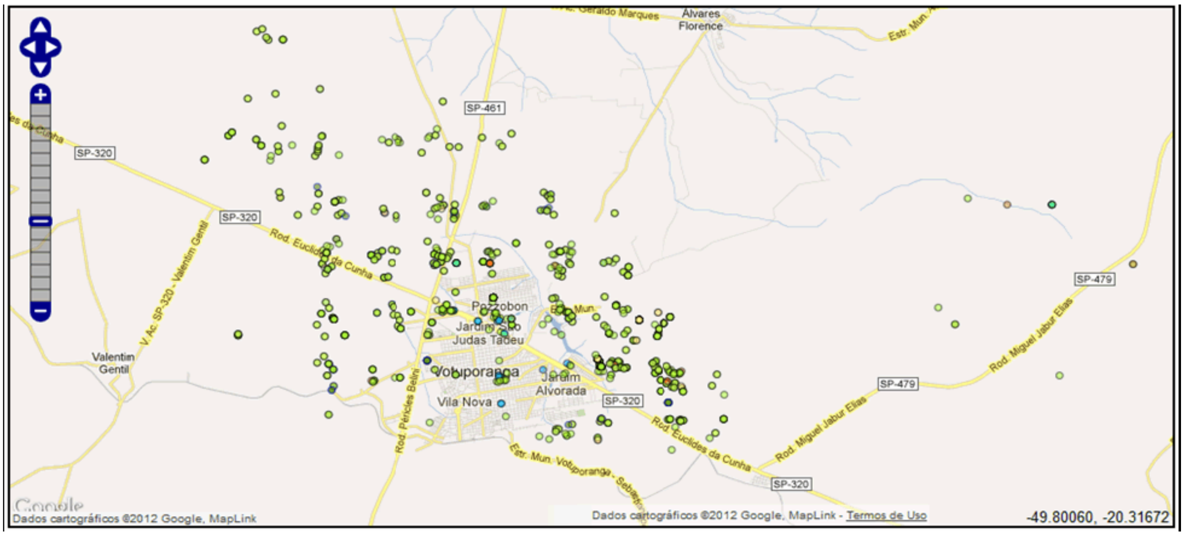

Fig. 1 Water resource distribution around the municipality of Votuporanga, in northwest of São Paulo State—Brazil.

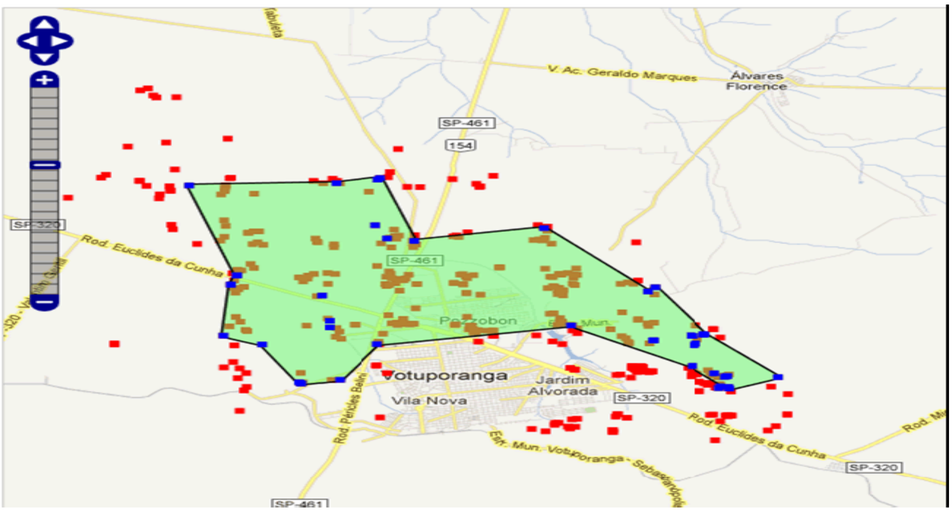

Fig. 2 First result: a cluster with 48 water resource catchment locations.

The second resulted cluster in this case is presented in Fig. 3, which groups six water resource catchment locations in a region of Votuporanga's urban area. The relevance of this result is that all of the grouped objects—blue squares—are quite close to each other, although being a small number, these water resource have been explored by administrators with a high level of education and under permission granted from the government department responsible for the regulation of natural resource exploration.

The last cluster resulting from a spatial data mining application is shown in Fig. 4. It is composed of 22 water resource catchment locations that are near downtown Votuporanga. The identified characteristics in this case are the exploration of these water resources executed by people with a high level of education and under a granted concession for their use. It means that the exploration there is legal and, according to the data collected on site, is a renewing process for the continuation of activities in which these resources are involved. Considering that the entire database has only 50 water resources administrated by people with this schooling—equivalent to 9.8%—the cluster in Fig. 4 represents 44% of this set, therefore, this case presents a high concentration in a small part of the analysed region.

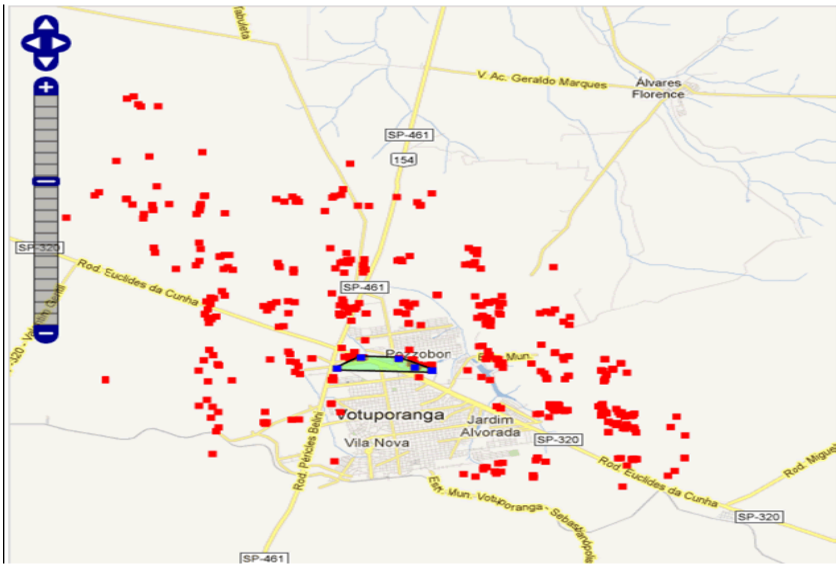

Fig. 3 Second result: a cluster with six water resource catchment locations.

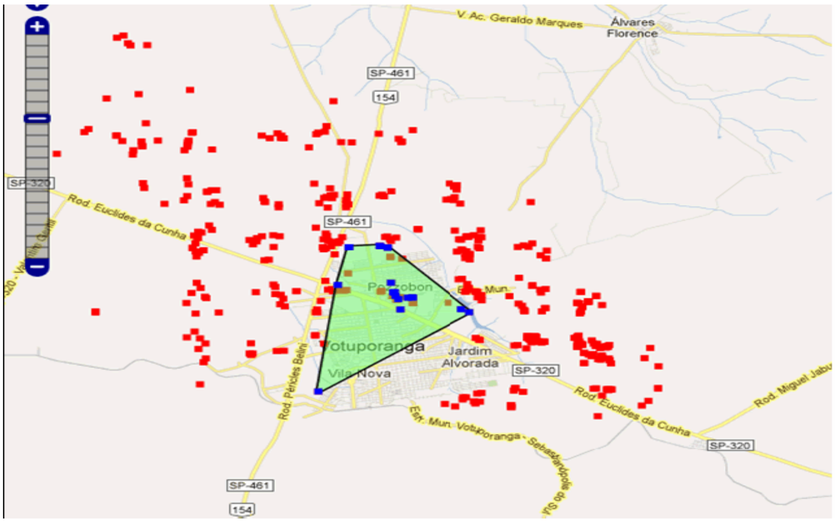

Fig. 4 Third result: a cluster with 22 water resource catchment locations.

Some important elements must be enunciated in order to demonstrate the contributions that spatial data mining provides for a more effective environmental management. The analysed database is a result of a project in execution since 2010, in which fieldwork was done in a region of São Paulo State and data about water resources were collected to compose a database used and managed by a GIS-based web information system. This technology was developed to aid environmental management based on information and advanced resources which enables the visualisation of water resource distribution around the focused area, the analysis of exploration and the condition of natural resources.

In addition to the support offered by the web information system, spatial data mining could return some information in which spatial location is considered for analysis and can be related to characteristics that describe the water resource conditions in database. The characteristics of water resource catchment presented in clusters shown in Figs. 2, 3 and 4 probably would not be identified by a conventional data mining, since the characteristics do not have predominant patterns. The fact that the objects—water resource catchments—present similar non-spatial characteristics and have their location near each other becomes the relevant characteristics for further analysis, resulting in information as presented previously. The information obtained from spatial data mining considering spatial location and non-spatial characteristics is quite useful for specialists in order to analyse them with another known information related to the spatial location, resulting in a rich support to management in decision making. One example of this is the case illustrated in Fig. 2, which contemplates a cluster of water resources under responsibility of administrators with low scholarity. This situation can be related to the fact that the focused area is undergoing a growing urbanisation process and, as a consequence, exploration of natural resources is on an unsustainable accelerated growth, which becomes worse when executed by managers with low scholarity—confirmed as the major occurrences on the visits done in fieldwork around Votuporanga's countryside.

The occurrences of these characteristics in addition to known information about the area indicate the need for training courses, conferences and education initiatives organised by professionals so as to contribute towards an improvement of this scenario, since increasing urbanisation allied to lack of knowledge and conscience results in unsustainable exploration.

Finally, some benefits that can be achieved by analysing spatial data mining results are:

(1) The identification of potential areas where exploration tends to become unsustainable or where it is not regularised is important so that prevention and corrective policies may be elaborated in time to control exploration and avoid significant losses;

(2) Environmental management can survey water consumption according to the administration profile or intended use of it, since in the analysed area the explored water is being used for animal and human consumption and the location of water catchment may bring about different policies according to administration profiles;

(3) The elements or activities that characterise the explored region can be related to some indicative returned by spatial data mining—such as increasing urbanisation and emerging recreation farms around urban area, which have driven daily drilling for water wells;

(4) Finally, some previously unknown situations and trends can be highlighted and aid in decision making, as the confirmation of the relationship between water exploration regularisation and scholarship of those who are responsible for it, since some of the water resources administrated by people with a higher scholarity level were under a license renewing process in order to continue the exploration. It is important to highlight that most of the water wells in urban areas are administrated by public government, which can justify the concern for regularisation.

4. Conclusion

Several technologies have been developed in order to support activities and aid decision making, which have provided information to enrich the policies and actions concept by environmental management. Besides these resources, additional and useful information can be obtained using other technologies for data analysing. One of them is spatial data mining, focused in this paper to illustrate a case study in Brazil. Through algorithms for spatial data mining, some information could be explicited besides the support already offered with the use of information systems for decision making. This contributes to increase the benefits obtained from better managed financial investments and/or concept of policies.

Although the case study related to water resource has a data set of only 507 geo-referenced locations which were explored, it was already possible to see the support which spatial data mining applied to environmental databases can provide to decision making, using similar spatial and non-spatial characteristics identified on data analysis to enrich the set of information provided by an information system. In addition to this, the use of spatial-temporal data can provide more information useful to establish a prediction of behaviour.

Finally, spatial data mining is useful for other scopes related to ecological issues, like global warming, genomics, animals, urban forestry control and marine application. As demonstrated, there are several databases supporting researches by organising and centralising data repositories so that information extraction from them is quite valued. Therefore, spatial data mining is a promising research area in which efforts have been applied seeking for improvements in algorithms, in order to obtain better performance and quality of results, since the range of issues for information about the behaviour and its correlations is wide.

Acknowledgments

The authors thank to CNPq for the financial support.

References

[1] Goodchild, M. F. 2007. “Citizens as Sensors: The World of Volunteered Geography.” *GeoJournal* 69 (4): 211-21.

[2] Wilson, J. P., and Fotheringham, A. S. 2008. “Geographic Information Science: An Introduction.” *The Handbook of Geographic Information Science*. Oxford: Blackwell, 1-10.

[3] Boyd, D. S., and Foody, G. M. 2011. “An Overview of Recent Remote Sensing and GIS Based Research in Ecological Informatics.” *Ecol. Inform*. 6 (1): 25-36.

[4] Van Aardt, J. A. N., and Wynne, R. H. 2007. “Examining Pine Spectral Separability Using Hyperspectral Data from an Airborne Sensor: An Extension of Field - Based Results.” *Int. J. Remote Sens.* 28 (2): 431-36.

[5] Wang, L., Sousa, W. P., Gong, P., and Biging, G. S. 2004. “Comparison of IKONOS and QuickBird Images for Mapping Mangrove Species on the Caribbean Coast of Panama.” *Remote Sens. Environ.* 91 (3-4): 432-40.

[6] Beeri, O., and Peled, A. 2009. “Geographical Model for Precise Agriculture Monitoring with Real-Time Remote Sensing.” *ISPRS J. Photogramm. Remote Sens.* 64 (1): 47-54.

[7] Falkenberg, N. R., Piccinni, G., Cothren, J. T., Leskovar, D. I., and Rush, C. M. 2007. “Remote Sensing of Biotic and Abiotic Stress for Irrigation Management of Cotton.” *Agric. Water Manag.* 87 (1): 23-31.

[8] Cammalleri, C., Andersonb, M. C., Ciraoloa, G., D'Ursoc, G., Kustasb, W. P., La Loggiaa, G., Minacapillid, M. 2012. “Applications of a Remote Sensing-Based Two-Source Energy Balance Algorithm for Mapping Surface Fluxes without in Situ Air Temperature Observations.” *Remote Sens. Environ.* 124: 502-15.

[9] Mennis, J., and Guo, D. 2009. “Spatial Data Mining and Geographic Knowledge Discovery—An Introduction.” *Comput. Environ. Urban Syst.* 33 (6): 403-8.

[10] Arentze, T. A. 2009. “Spatial Data Mining, Cluster and Pattern Recognition.” In *Int. Encycl. Hum. Geogr.* Oxford: Elsevier, 325-31.

[11] Jin, H., and Miao, B. 2010. “The Research Progress of Spatial Data Mining Technique.” *In Proceeding of 2010 3rd International Conference on Computer Science and Information Technology*, 81-4.

[12] Shekhar, S., Zhang, P., and Huang, Y. 2005. “Spatial Data Mining.” *Data Min. Knowl. Discov. Handb.*, Springer, US. 833-51.

[13] Shekhar, S., Evans, M. R., Kang, J. M., and Mohan, P. 2011. “Identifying Patterns in Spatial Information: A Survey of Methods.” *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* 1 (3): 193-214.

[14] Koperski, K., and Han, J. 1995. “Discovery of Spatial Association Rules in Geographic Information Databases.” In *Advances in Spatial Databases,* Portland: Springer Berlin Heidelberg, 47-66.

[15] Koperski, K., Han, J., and Stefanovic, N. 1998. “An Efficient Two-Step Method for Classification of Spatial Data.” *In Proceeding of 1998 International Symposium on Spatial Data Handling SDH’98*, 45-54.

[16] Li, G., Peng, R., Zheng, Y., and Zhao, J. 2010. “Spatial Data Mining and Its Application in Marine Geographical Information System.” *In Proceeding of The 2nd Conference on Environmental Science and Information Application Technology*, 514-16.

[17] Zahiduzzaman, A. K. M., Khan, M., and Rahman, R. M. 2010. “Spatial Data Mining on Literacy Rates and Educational Establishments in Bangladesh.” *In 13th International Conference on Computer and Information Technology (ICCIT),* 394-99.

[18] Birant, D., and Kut, A. 2007. “ST-DBSCAN: An Algorithm for Clustering Spatial-Temporal Data.” *Data Knowl. Eng.* 60 (1): 208-21.

[19] Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise.” *In Proceeding of the*

Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 226-31.

[20] Liu, P., Zhou, D., and Wu, N. 2007. "VDBSCAN: Varied Density Based Spatial Clustering of Applications with Noise." In *Proceeding of the Int. Conf. Serv. Syst. Serv. Manag.,* 1-4.

[21] Ng, R. T. 2002. "CLARANS: A Method for Clustering Objects for Spatial Data Mining." *IEEE Trans. Knowl. Data Eng.* 14 (5): 1003-16.

[22] Ichiba, F. T. 2012. "Multi-relational Prospecting Algorithm for Spatial Data." Postgraduate in Computer Science, São Paulo State University, São José do Rio Preto. (in Portuguese).

[23] Wang, J., Chen, X., Zhou, K., Zhang, H., and Wang, W. 2009. "Research of GIS-Based Spatial Data Mining Model." *In Proceeding of the Second International Workshop on Knowledge Discovery and Data Mining*, 159-62.

[24] Valêncio, C. R., Carvalho, A. C., Jardini, T., Scarpelini Neto, P., Ichiba, T. F., and Medeiros, C. A. 2011. "Water Resources Management Supported by Geographic Information Systems." In *Proceeding of XIX Simpósio Brasileiro de Recursos Hídricos,* 224.

[25] Valêncio, C. R., Carvalho, A. C., Jardini, T., Ichiba, T. F., Scarpelini Neto, P., and Laurenti, C. H. 2010. "Georeferenced Computational System for Maintenance of Water Resource's Users." In *IADIS Ibero-Americana WWW/Internet 2010*, 364-68.