International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 20 (2017) pp. 10057-10062

DATA MINING: An environmental quality analysis in the Negro River Basin (Colombia)

López Sánchez, Wilson Ricardo[1]; Rodríguez Miranda, Juan Pablo[2]; García Ubaque and Cesar Augusto[3]

[1] Electronic Engineer. (Candidate) Magister in Information and Telecommunications Sciences. LASER Research Group. Francisco Jose de Caldas District University, Colombia.

[2] Sanitary and Environmental Engineering. Magister in Environmental Engineering. PhD (Candidate) Associate Professor. Universidad Distrital Francisco Jose de Caldas. Director of the AQUAFORMAT research group, Colombia.

[3] Civil Engineer. Doctor of Engineering. Associate professor. Francisco Jose de Caldas District University. Director of GIICUD research group, Colombia.

Abstract

This paper presents the assessment of the concurrent environmental quality evaluated in the conditions in the Watershed of the Negro River, integrating the variables of the water quality (BOD, TSS, N-NO_2 y P_{total}) and precipitation. Using the data mining technique establishes membership functions and describes data and analyzed variables. In the body of surface water, there is a detriment of the environmental quality in the medium and low basin, which is an obvious contamination of the river and therefore the need to do short-term intervention actions in the basin.

Keywords: Watershed, data mining, environmental quality.

INTRODUCTION

The applications of computational systems for making decision and prediction of the behavior of natural phenomena have been increased in terms of the techniques that may represent the conditions and abstraction of the phenomenon (Refonaa J, 2015). The information obtained from different natural phenomena have been used in different computational techniques such as machine learning, likewise databases are widely used to find important information in processes known as data mining (Pulvirenti, 2014; Karim, 2016).

Data mining or knowledge discovery in databases, consists to extract information from the data, to give it meaning and to draw useful conclusions from it, by describing patterns in large data sets provided for finding intelligible models from them (Benítez, 2013; Medina, 2014; Escobar H. , 2016; García M. , 2007). It provides a response according to the linguistic and verbal information of data, considering the assignment of the partial belongings of any object to different subsets of a universal set, instead of belonging to a single set and this membership function assume values between zero and one. The purposes of this technique, consist in the prediction through the classification (associated a discrete value and the objective is to maximize the predictive power of the classification), regression (it has associated a real value and the objective is to learn a real function, with that can minimize the error between the predicted value and the actual value) through a set of input and output attributes, the value of which can be a category or numerical value, i.e. predict the output value; additionally another purpose , is the description through grouping (to obtain groups of natural form when applying criteria of similarity data), this one presented without labeling or enumerate, which only possess attributes of entrance and the objective is to describe data (García M. , 2007; Riquelme, 2006; Ruiz, 2006; Itati, 2012).

Among the different fields, the search for missing parameters and estimation of parameters is considered (Ssali, 2008). This computational technique can cover several areas of knowledge where one has a way of acquiring data or a determined database to which can be made studies of different types (Zhun, 2016) with the aim of obtaining a relation or prediction of one or various variables of the data with which it is counted. Many models describe the behavior of different physical phenomena that require complicated calculations and are not adaptive models (Chapra, 1987; Chapra S, 2008). However, with data mining, relevant information can be obtained to estimate missing data and, of course, to approximate the knowledge and behavior of the analyzed natural phenomena.

This technique, is a method of approximation where there are no mathematical equations, however the uncertainties and complications of the model are included in the procedure of descriptive diffuse inference (Erkan, 2009). The applications of techniques are usually in the modeling of surface and ground water quality, estimation of water quality through satellite imagery, prediction of earthquakes, prediction of basin levels (Bonansea, 2015; Harvey, 2015); recognition of water quality patterns and sustainable use of water,

identification of ecosystem functioning models, improvement of management and control of wastewater treatment plants, urban planning (Ay, 2014; Sari, 2013; Pai, 2011; Ross, 2010).

This paper presents the analysis of the variables of precipitation and water quality (BOD, TSS, N -NO_2 y P_{total}) in order to understand the patterns of behavior, extract attributes, consider membership functions and describe significant data in the watershed of the Negro River (Colombia).

MATERIALS AND METHODS

The method used is a combination between the real and exact observation and the knowledge of an empirical, complex situation and inductive reasoning, which would consist in deriving a new knowledge from particular phenomena and knowledge already obtained, and establishing propositions analyzed from their causes and real effects, that is, from the particular to the general (Vergel, 2010; Balestrini, 2001). It is worth mentioning that according to the analysis and scope of the results, the type of research is analytical - quasi experimental, since it analyzes an event and understands it, in terms of its obvious aspects and discovers the elements that make up the totality and the connections that explain its integration, that is, it facilitates the study and deeper understanding of the event under study (Hurtado J., 2000; Vergel G., 2010; Hernández, 2010).

Precipitation information was obtained from the climatological stations of the Cundinamarca´s Autonomous Regional Corporation (CAR) located in each of the municipalities belonging to the watershed of the Negro River; information of the water quality parameters BOD, TSS, N - NO_2 and P_{total}, related to the surface water quality as the wastewater treatment plants (including treatment flow) located in towns in to the basin in question, were taken from the Cundinamarca´s Environmental Laboratory of the Regional Autonomous Corporation (CAR).

The analysis period for precipitation and water quality information is from year 2012 to 2014. A database was developed with the estimation or replacement of missing data, thus a database of 155 is constructed with 5 different variables and for this the mean and variance for each analyzed variable is determined, then it is ordered upwards with respect to the calculated variance.

RESULTS

The following are the results of applying data mining in the Negro River Basin.

In Figure 1, for the BOD parameter, is observed a density and reciprocity of the connecting line entities between the stations and the analysis period (2012-2014), which shows a good environmental quality in the upper and lower basins and which establishes a reduced variability, higher stability of the phenomenon and a homogeneity behavior of the environmental quality, having a marked segment of data in the discharge of the municipality of the Vega that leads to a detriment of the environmental quality of the surface water.

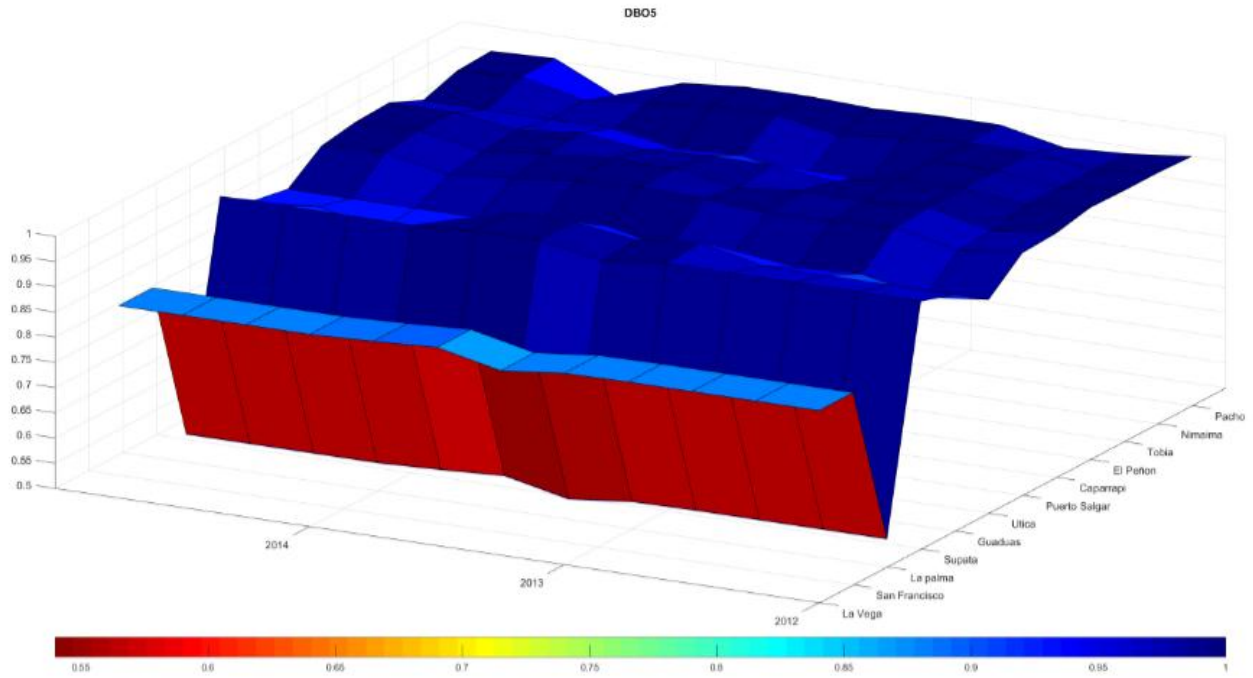

Figure 1. Three-dimensional diagram for BOD.

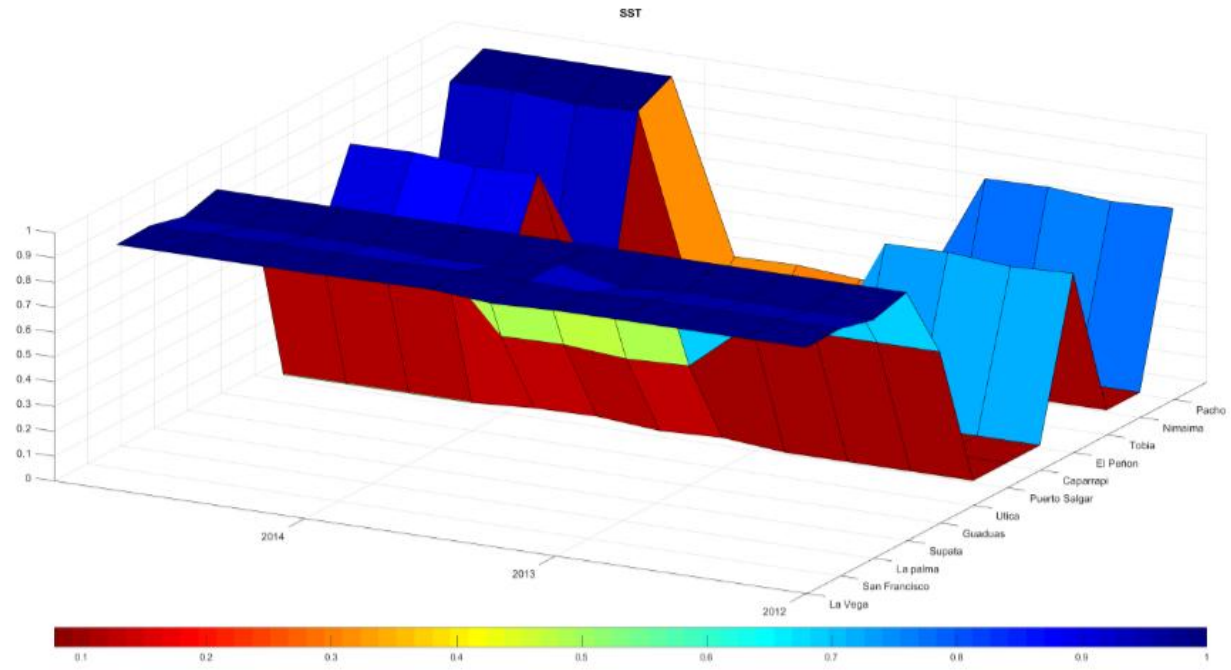

Figure 2. Three-dimensional diagram for TSS.

In Figure 2, for the TSS parameter, there is an identifiable behavior of patterns of high variability of environmental quality in the period 2012 to 2014, especially the impact of spills of municipalities with more population that contribute a considerable detriment of the environmental quality to the body of water. In the middle and lower basin, a decrease in environmental quality is observed, due to a structure caused by the high variability of the reported information and the high dispersion with recurrent frequency.

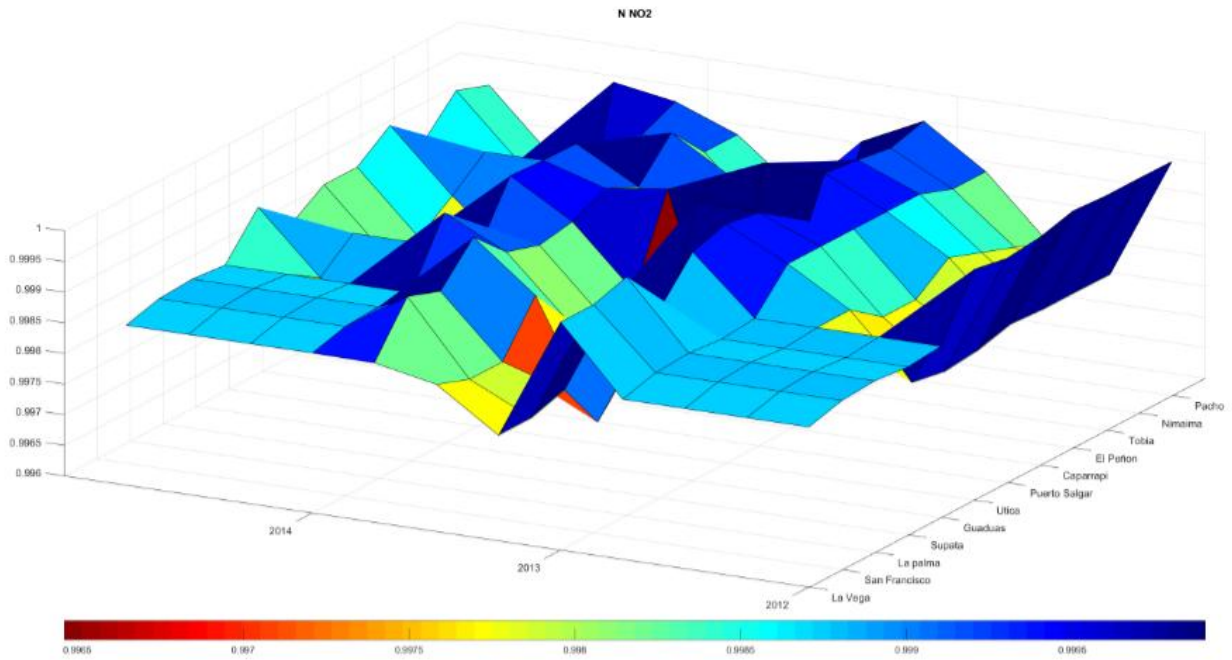

Figure 3. Three - dimensional diagram for N-NO_2.

In Figure 3, the behavior of the parameter N- NO_2 is observed, with a correlation pattern variability and heterogeneity in environmental quality between good and fair, confluent in a topology environmental quality in the basin in the period 2012 to 2014, although in the period 2013, the behavior is of considerable traffic with segments between line entities between data and data of environmental quality, being evident a gradual detriment of the environmental quality.

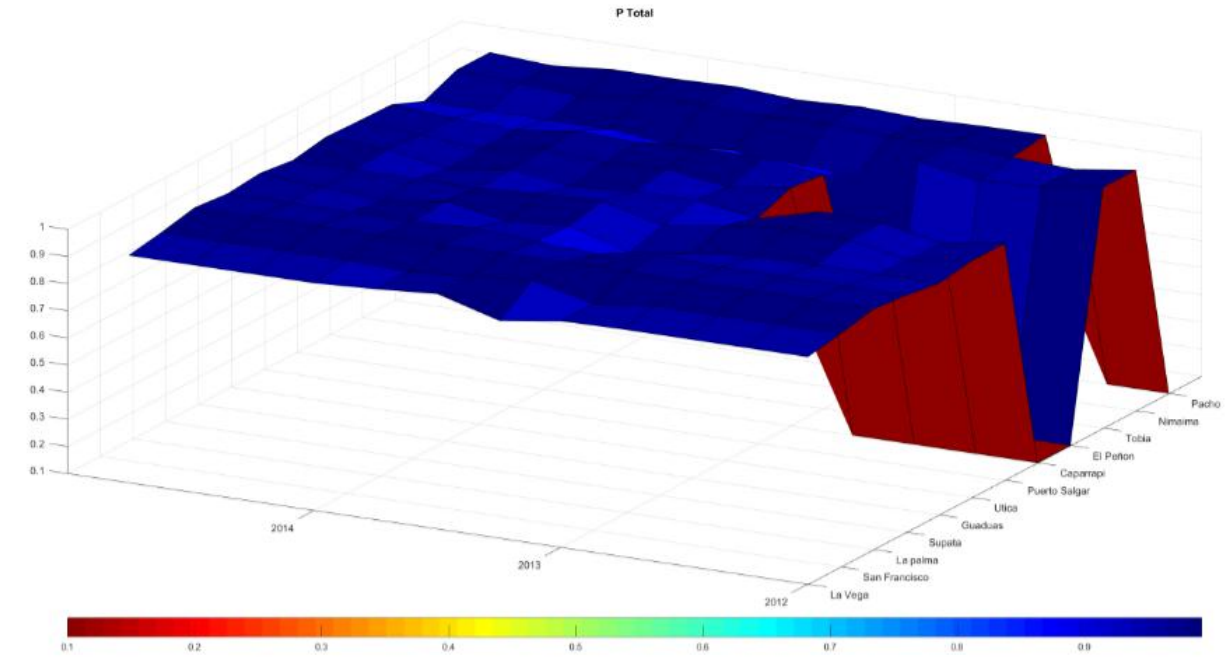

Figure 4. Three-dimensional Diagram for P_{total}.

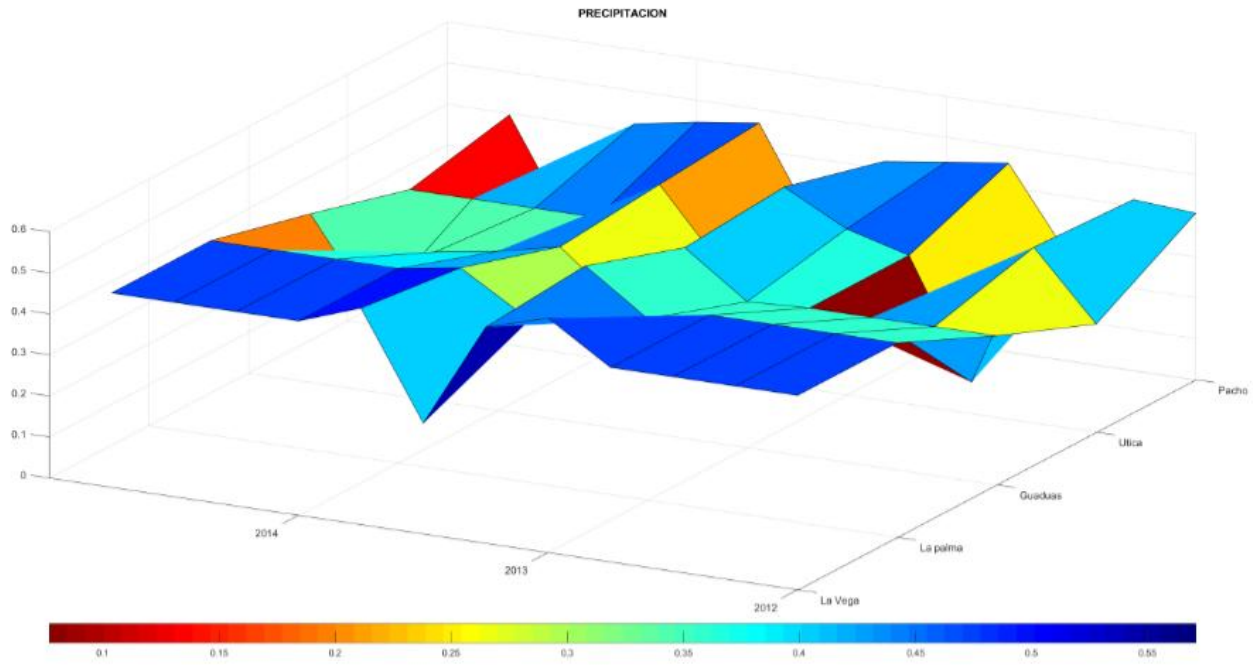

Figure 5. Three-dimensional diagram for precipitation.

In Figure 4, for the P_{Total} parameter, it is noted connecting line entities between stations and analyzed period 2012-2014, with low temporal correlation, reduced variability and marked homogeneity in the analyzed phenomenon, however, it is noted that in Puerto Salgar, Anapoima and Guadas municipalities, a discharge of waste water with high content of nitrogen, generates a detriment of the environmental quality.

Figure 5 shows a behavior of medium stability, dispersion and marked heterogeneity for the connection line entities between stations and in the analyzed period for the precipitation variable in the Negro River Basin, indicating a considerable amount of water that enters the basin and therefore a relative dilution of the conservative contaminants.

CONCLUSIONS

Data mining applied to the analysis of the variables of water quality and precipitation in the Negro River basin, considers understand and comprehend patterns of behavior, extract attributes, consider membership functions and describe significant data of BOD, TSS, N -NO_2, P_{total} and precipitation, in order to identify suitable actions in the short term intervention in the basin in terms of spatial identification of scenarios in sections or sectors of the surface water bodies especially in the middle and lower basin, due to the high anthropic pressure, the predominant environmental effect and the evident alteration of the environmental quality in the river. With the membership functions, it is possible to establish a marginal approach to the installation or optimization of wastewater treatment systems (WWTS / WWTP), related to aspects of the sensitivity, adaptability of the technology to be implemented for the increase of quality in the middle and lower basin.

ACKNOWLEDGEMENTS

Authors grateful to the Engineering Doctorate program of the Francisco Jose de Caldas District University (Bogotá, Colombia), the Cundinamarca`s Autonomous Regional Corporation (CAR) and the Hydrology, Meteorology and Environmental Studies Institute (IDEAM) for the information provided of the analyzed Basin.

REFERENCES

[1] Ay, M. (2014). Modelling of chemical oxygen demand by using ANNs, ANFIS and k-means clustering techniques. *Journal of Hydrology*, Numero 511, pp 279 - 289.

[2] Balestrini, M. (2001). *Cómo se elabora el proyecto de investigación.* Caracas, Venezuela: BL Consultores asociados.

[3] Benítez, R. (2013). *Inteligencia artificial avanzada.* España: UOC. Fundación Universidad Oberta de Cataluña

[4] Bonansea, M. (2015). Using multi-temporal Landsat imagery and linear mixed models for assessing water quality parameters in Río Tercero reservoir (Argentina). *Remote Sensing of Environment*, Vol 158, No 1, March, 28 -41. .

[5] Chapra S. (2008). QUAL 2K: A Modeling Framework for Simulating River and Stream Water Quality. En Chapra S, *QUAL 2K: A Modeling Framework for Simulating River and Stream Water Quality.* (pág. 89). USA: EPA. Mc Graw Hill.

[6] Chapra, S. (1987). Surface Water Quality Modelling. En Chapra S, *The Enhanced Stream Water Quality Models QUAL2E and QUAL2E-UNCAS, EPA/600/3-87- 007,* (pág. 189). USA: Mc Graw Hill. Brown, L.C., and Barnwell, T.O. Environmental Protection Agency,.

[7] Erkan, M. (2009). River flow estimation from upstream flow records by artificial intelligence methods. *Journal of Hydrology*, Numero 369, pp 71 - 77

[8] Escobar, H. (2016). Aplicaciones de minería de datos en marketing. . *Publicando*, Volumen 3, Numero 8, pp 503-512.

[9] Escobar, M. (2016). DIseño de un sistema experto para la reutilización de aguas residuales. *Ciencia e ingenieria neogranadina*, Volumen 26, numero 2, pp 21- 34.

[10] García, M. (2007). *Aplicación de técnicas metaheurísticas en minería de datos.* España: Universidad de Laguna. Servicio de publicaciones. Ciencias y tecnologías.

[11] Harvey, T. (2015). Satellite-based water quality monitoring for improved spatial and temporal retrieval of chlorophyll-a in coastal waters. *Remote Sensing of Environment*, Vol 158, No 1, March, 417-430.

[12] Hernández, R. (2010). *Metodología de la investigación.* . México: Mc Graw Hill.

[13] Hurtado J. (2000). *Metodología de la investigación holística.* . Caracas: Fundación SYPAL.

[14] Itati, M. (2012). Revisión de algoritmos de Redes Neuronales en dos herramientas de Minería de Datos. *Técnica administrattiva*, Volumen 11, numero 4, pp 10-15.

[15] Karim, M. (2016). A comprehensive study on the effects of using data mining techniques to predict tie

strength, . *Computers in Human Behavior*, Vol 60, Julio. 534 - 541.

[16] Medina, F. (2014). Funcionalidades de la minería de datos. . *Revista Ingeniería y región*, Volumen 12, Numero Noviembre, pp 31 - 40.

[17] Pai, T. (2011). Predicting effluent from the wastewater treatment plant of industrial park based on fuzzy network and influent quality. *Applied Mathematical Modelling*, Numero 35, pp 3674 -3684

[18] Pulvirenti, L. (2014). "La discriminación de las superficies de agua, las lluvias abundantes, y Wet nieve usando las observaciones COSMO-SkyMed de eventos de tiempo severo",. *IEEE Transactions on ciencias de la tierra y la teledetección*, Vol 52, Numero 2, 152 - 189.

[19] Refonaa J. (2015). "Analysis and prediction of natural disaster using spatial data mining technique" . *International Conference on Circuit, Power and Computing Technologies*, (págs. 1-10). USA.

[20] Riquelme, J. (2006). Minería de datos: conceptor y tendencias. *Revista Iberoamericana de Inteligencia Artificial*, Volumen 29, pp 11-18.

[21] Ross, T. (2010). *Fuzzy Logic, with engineerin applications.* New mexico, USA: WILEY. Third edition.

[22] Ruiz, R. (2006). Presentación: minería de datos. *Revista Iberoamericana de Inteligencia Artificial*, Numero 29, pp 7-9.

[23] Sari, H. (2013). Fuzzy-logic modeling of Fenton's strong chemical oxidation process treating three types of landfill leachates. *Environ. Sci. Pollut*, Numero 20, pp 4235-4253.

[24] Ssali, G. (2008). Computational intelligence and decision trees for missing data estimation. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, (págs. 20-25). USA.

[25] Vergel G. (2010). *Metodología. Un manual para la elaboración de diseños y proyectos de investigación. Compilación y ampliación temática.* Barranquilla: Publicaciones Corporación UNICOSTA.

[26] Zhun, J. (2016). Advances and challenges in building engineering and data mining applications for energy-efficient communities. *Sustainable Cities and Society,* , Vol 25, August, 33 - 38.