Journal of Water Resource and Protection, 2016, 8, 522-533
Published Online April 2016 in SciRes. http://www.scirp.org/journal/jwarp
http://dx.doi.org/10.4236/jwarp.2016.84044

Data Mining of Historic Hydrogeological and Socioeconomic Data Bases of the Toluca Valley, Mexico

Oliver López-Corona[1,2], Oscar Escolero Fuentes[3], Eric Morales-Casique[3], Pablo Padilla Longoria[4], Tomás González Moran[5]

[1]Instituto de Investigación sobre Desarrollo Sustentable y Equidad Social, Universidad Iberoamericana, Ciudad de México, México
[2]Centro de Ciencias de la Complejidad (C3), Universidad Nacional Autónoma de México, Mexico City, Mexico
[3]Instituto de Geología, Universidad Nacional Autónoma de México, Ciudad de México, México
[4]Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Ciudad de México, México
[5]Instituto de Geofísica, Universidad Nacional Autónoma de México, Ciudad de México, México
Email: lopezoliverx@otrasenda.org

Received 1 December 2015; accepted 26 April 2016; published 29 April 2016

Abstract

In this paper we used several data mining techniques to analyze the coevolution of hydrogeological and socioeconomical data for the Toluca Valley in Mexico. We found non trivial relations between two historic data bases that make clear that groundwater and economy may be much more linked than it was thought before. In particular, we found that hydrogeological data trends change during economical crisis and election years in Mexico. This shows that different macroeconomical policies implemented by several administrations have a direct impact in the way groundwater is used. We also found that hydrogoelogical data evolve in the direction of population transformation from rural to urban, which could represent a whole paradigm shift in groundwater management with profound repercussions in policy making.

Keywords

Data Mining, Economy, Data Science, Groundwater

1. Introduction

Recently, there has been important progress in techniques for observing the environment, and as a result, the

How to cite this paper: López-Corona, O., Fuentes, O.E., Morales-Casique, E., Longoria, P.P. and Moran, T.G. (2016) Data Mining of Historic Hydrogeological and Socioeconomic Data Bases of the Toluca Valley, Mexico. *Journal of Water Resource and Protection*, **8**, 522-533. http://dx.doi.org/10.4236/jwarp.2016.84044

volume of remote sensing data such as satellite data and point data acquired from various kinds of land observation instruments, has increased. Lots of useful and detailed data, which were very difficult to get before, are now available, and this contributes to the progress of various research fields. However, because of the sheer volume of new data, researchers have not been able to process it using traditional analytical techniques. Much of the information collected in the past few years is simply being stored [1].

"If you torture the data long enough, nature will confess," said the economist, 1991 Nobel laureate Ronald Coase. Although this statement is still true, this objective is not easy. First, "enough time" can be in practice "too long" in many applications and therefore unacceptable. Second, to obtain "confessions" from large data sets we must use the state of the art tools in "torture". Third, as nature has confessed some of her most hidden secrets, it seems that she has become more stubborn and unwilling to share any more [2].

In our view, much of the best techniques of "torture" can be found in what is now known as data mining. Data mining is the essential ingredient in the more general process of knowledge discovery in databases (KDD). The idea is that by automatically sifting large amounts of data should be possible to extract nontrivial knowledge inherent in them. Data mining has become fashionable, not only in computer science, but especially in business, but why is it different from statistics? Certainly data mining uses statistics and is even based on it, but it also incorporates database management techniques and modern artificial intelligence algorithms. It includes all this, but it is different at the same time. More than a collection of many types of analysis, data mining is distinguished by a distinctive approach, a new attitude toward data analysis. The emphasis is not so much in the extraction of facts, but in the generation of hypotheses. Its aim has more to do with generating new and better questions than to refine answers to traditional ones. To achieve this, data mining uses a vast collection of statistical techniques and artificial intelligence methods such as: neural networks, factor analysis, time series analysis, Bayesian networks, decision trees, statistical models, multivariate statistical analysis and clustering analysis [3] [4].

Data mining has been applied with success in many areas of science such as Biology [5]-[8], Astronomy [9]-[13] and Medicine [14] [15] just to mention some. In particular, due to the multisystem nature of Earth sciences, we consider that the incorporation of data mining to this discipline would provide a new perspective to analyze old problems and would likely suggest new ones [16]-[21].

2. Site of Study

The Toluca Valley is located in the State of Mexico (**Figure 1**), within the watershed of Lerma in the South of

Figure 1. Localization map of the Toluca Valley in Mexico. The Toluca Valley is a valley in central Mexico, just west of Mexico City. Since the 1940s, there has been significant environmental degradation in the valley, with the loss of forests, soil erosion, falling water tables and water pollution due to growth in industry and population.

the Mexican Plateau. It is bounded from the North by the aquifer Atlacomulco Ixtlahuaca, from the South by the Tenango hill, from the South-West by the Nevado de Toluca volcano and from the East by the Sierra de las Cruces mountain chain; covering approximately 2738 km^2.

The Toluca Valley is part of the Rio Lerma basin which has a good potential for groundwater exploitation which in fact not only is used by local farmers from Toluca and other small cities, but it also exports large volumes through the Lerma's well battery system for Mexico City water supply, becoming a strategic source of water.

According to INEGI (National Institute of Statistics and Geography) census data, the population of the State of Mexico, is of 1,107,964 inhabitants, accounting for 13% of Mexico's population, being the entity with the highest population density. The state's population growth has been uneven, but has been localized on narrowly defined areas including the municipalities of Toluca, Metepec, Lerma, and Zinacantepec. From 1990 to 1995 the population grew nearly 17% due to rapid industrial growth and residential development, trend that continues to our days. It should be noted that between 1950 and the eighties, the State of Mexico moved from the seventh to the first place among the 32 states in terms of total population. Much of this increase, both state and regional, occurred during the decades of the 60's and 70's when the average annual growth rates were of 7% and 4% respectively [22] [23].

Outside the metropolitan area, the economy is still based on agriculture and livestock, with some income from tourism. Only a little over 4% of the total municipal population engages in agriculture raising corn, wheat, beans, potatoes, peas, fava beans and oats on a little over half of the municipality's territory. Livestock raising is a greater source of income with 10,286 sites producing cattle, porks, sheep and domestic fowl [24].

3. Method

In this work we used data-mining techniques to analyze a 40-year piezometric level data set from the Toluca Valley in Central Mexico. The monitoring network was built in the late 60's to register hydraulic head in the aquifer. Each monitoring location in the network consists of a nest of piezometers (bores) with up to eight piezometers installed at different depths ranging from 10 to 200 m (**Figure 1**). Hydraulic head has been measured in the network in a monthly basis since 1969 (**Figure 2**), it is currently operated by the National Water Comission (CONAGUA) and it provides information to analyze the space-time response of the hydrogeologic system to external forcing, among which is pumping [25].

To explore the relations between the evolution of the groundwater system and socioeconomic factors we selected seven socioeconomic variables from INEGI, Mexico's National Institute of Statistics and Geography. The

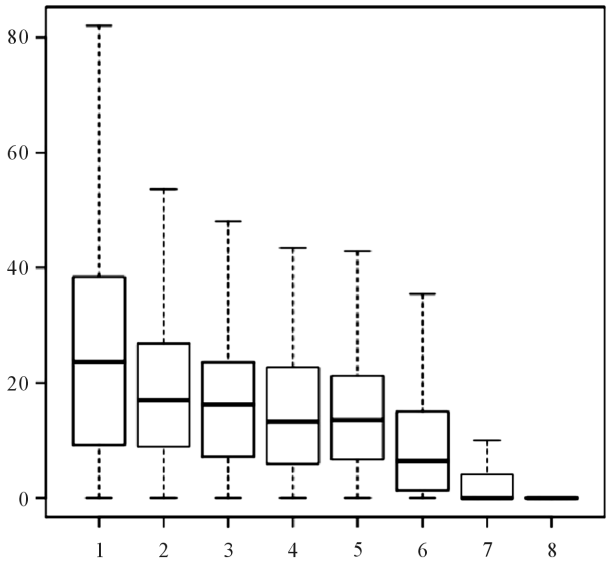

Figure 2. Boxplot graphs for the peizometric lectures [m] of the eight bores over all well and years. Bores are organized from the deepest to shallowest.

variables selected are: 1) gross national domestic product (GNP), 2) total country population, 3) urban country population, and 4) rural population in the country, 5) total estate population, 6) urban state population, and 7) rural state population.

Merging large data bases acquired from different sources, purposes and having different data representation has become such an important problem in data mining that data cleaning has been considered a crucial first step in knowledge discovery [26] [27]. The current work started with extensive data cleaning (original data set can be found in here: http://www.geologia-feflow.unam.mx/documentos/base_toluca_original.ods) that included removal of text and undesired headers, annotation standardization and search for missing and misleading data and data standardization. In general missing and misleading data was ignored. For this purpose we used awk scripts, SQL management and manual supervision. Then we realized exploratory statistical, clustering and variance analysis (ANOVA) to understand data inner dependence and structure. We merge both data bases in one, that we analyzed from a multivariate statistical point of view. Principal component and canonical correspondence analysis were performed for identify suitable variables for a classification and clustering analysis.

We consider relevant to point out that only free software was used for this work, from a Debian Squeeze Gnu-Linux operating system, awk for prepossessing; MySQL, OpenOffice and Gnumeric for data management; R for exploratory and multivariate analysis; Weka for classification (J48 algorithm) and clustering analysis (k-means) and LyX for paper text processing. This note is important since software licenses prices may in fact impose severe limitations to science in developing countries such as Mexico and more important is the part of the struggle to ensure free access of scientific products financed with government money.

4. Results

Exploratory statistical analysis showed in **Table 1** and **Figure 2**, **Figure 3** and **Figure 4** reveal that only years and piezometers variables have inner statistical independence and that socioeconomic variables are very high correlated (**Table 2**). Piezometers were grouped according to depth as: deep (bore 1), medium (bores 2 - 5) and shallow (bores 6 - 8). Similarly, years were grouped according to time periods: from 1969 to 1977; from 1978 to 1989; from 1990 to 1996 and from 1997 to 2002.

Table 1. Anova analysis results.

	F	P (<F)	Significance
Wells (w)	0.2537	0.6145	No
Years (y)	72.0601	2.2E−16	Yes
Bores (b)	4870.72	2.2E−16	Yes
w+y	0.3379	0.561	No
w+b	39.5954	3.2E−10	Yes
y+b	65.4259	6.5E−16	Yes
w+y+b	1.7819	0.1819	No

Table 2. Correlation matrix.

Correlations	MTP	MRP	MUP	STP	SRP	SUP	GNP
MTP[a]	1						
MRP	−0.975	1					
MUP	0.975	−0.998	1				
STP	0.995	−0.96	0.96	1			
STP	−0.924	0.89	−0.89	−0.89	1		
SUP	0.924	−0.89	0.89	0.89	−0.998	1	
GNP	0.986	−0.99	0.99	0.99	−0.99	0.99	1

[a]About the abbreviations: MTP is the Mexico's Total Population, MRP is the rural population and MUP the urban. For its part, STP is the State's Total Population, SRP the rural population and SUP the urban.

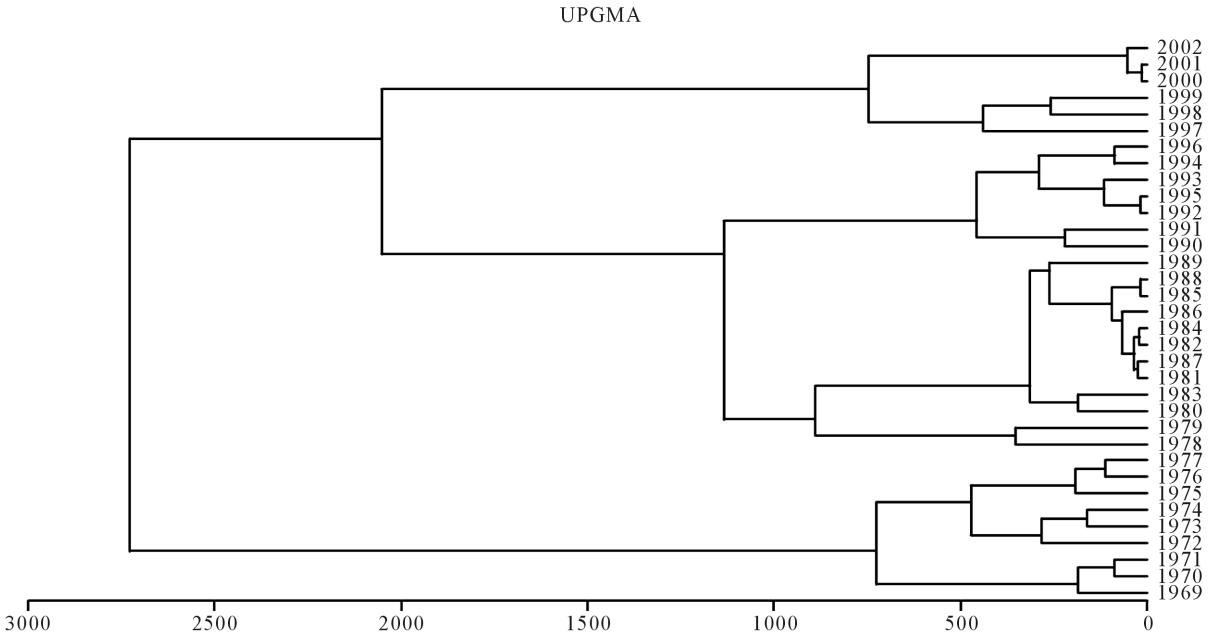

Figure 3. Years dendogram.

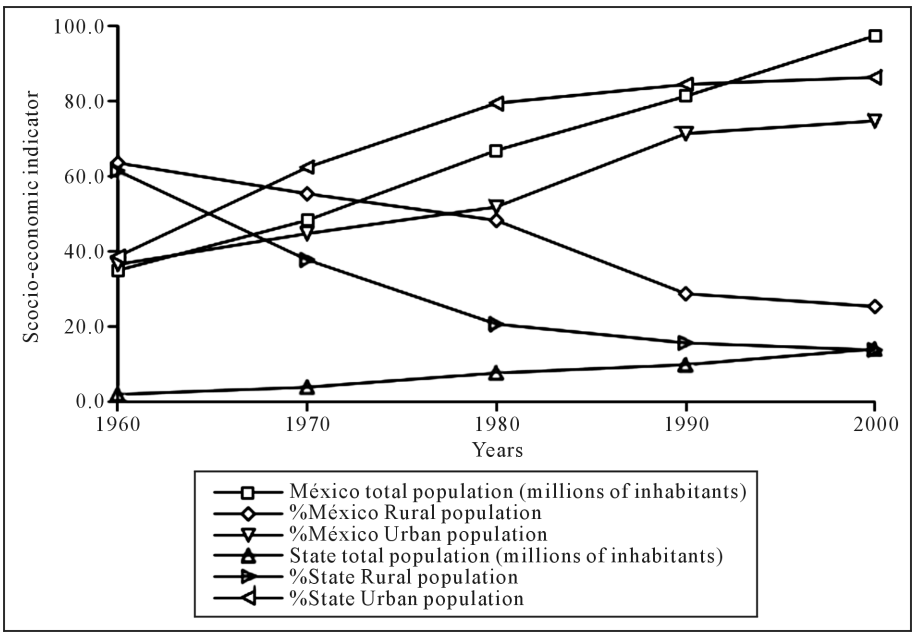

Figure 4. Time evolution of the different socioeconomic indicators considered in this work.

From multivariate analysis (**Figure 5**, **Figure 6**) we can observe that data variance is mostly explained by state population structure and GNP. In **Figure 6**, years are highlight in different color corresponding to decades and space was divided in four quadrant that correspond to different combinations of population structure and GNP increases. For clarity the labels were omitted in **Figure 5** in which time begins at the bottom-right quadrant IV, then follow a crescent behavior to quadrant I, decreases to IV crossing to III, increases again to II and finally decreases to III. The final state is characterized by a high urban population and decreasing GNP tendency (**Figures 7-9**). In general, data arrange themselves so that as time grows, they move from a population structure predominantly rural to a basically urban, experiencing GNP ups and downs that correspond to different periods of the country economy. It is even possible to identify the 1987 and 1994 crisis years.

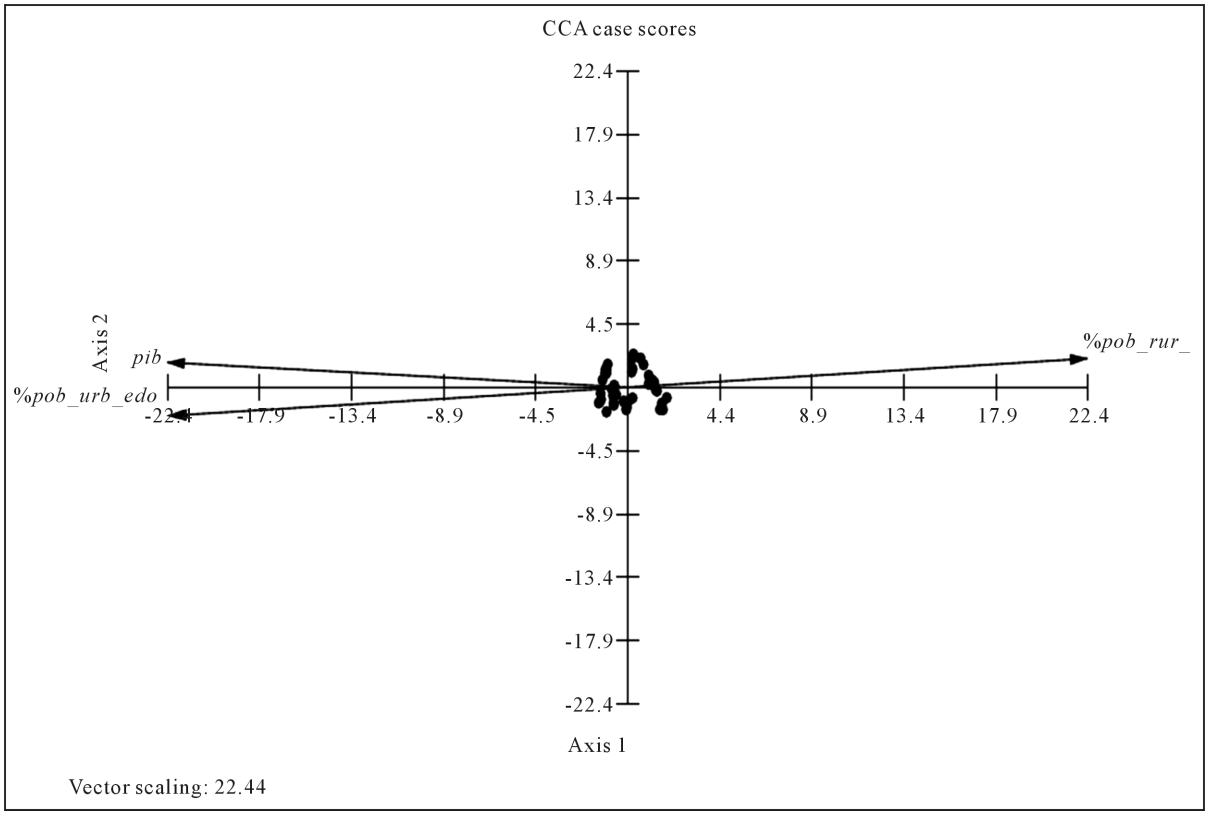

Figure 5. Canonical correspondence analysis for years with socioeconomics vectors.

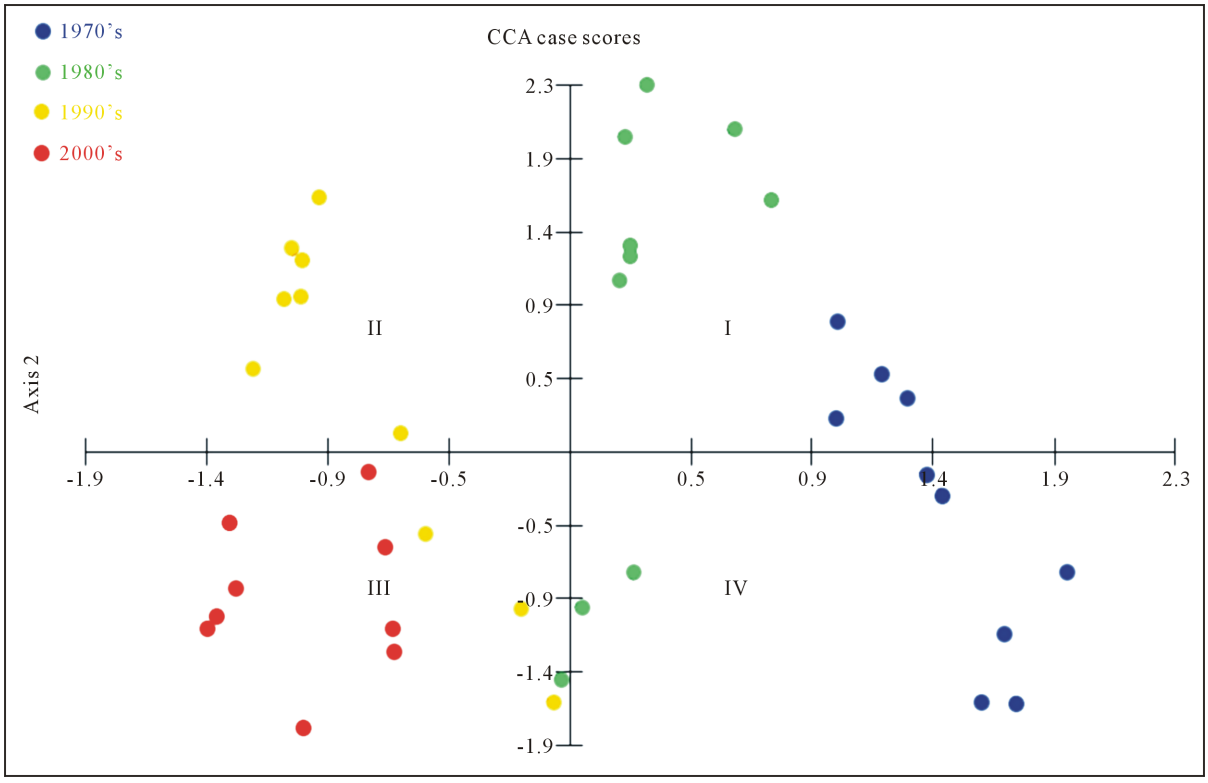

Figure 6. Canonical correspondence analysis for years without socio-economic vectors.

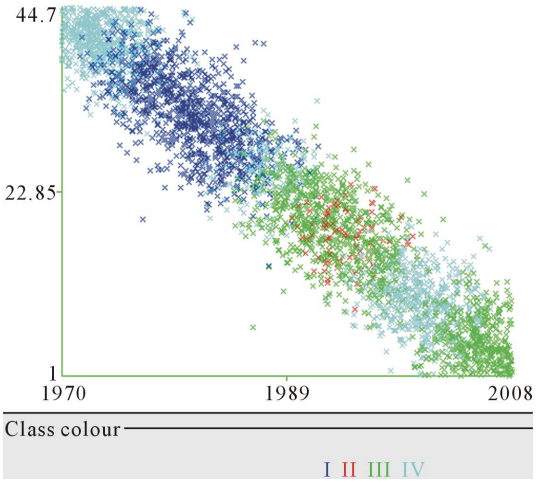

Figure 7. State rural population percentage vs time. Class colour correspond to CCA quadrant.

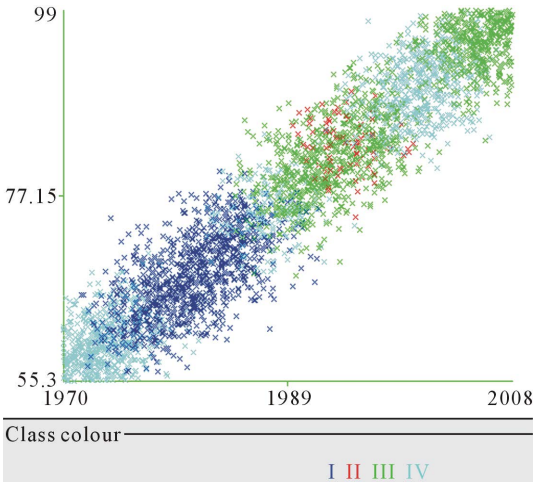

Figure 8. State urban population percentage Vs time. Class colour correspond to CCA quadrant.

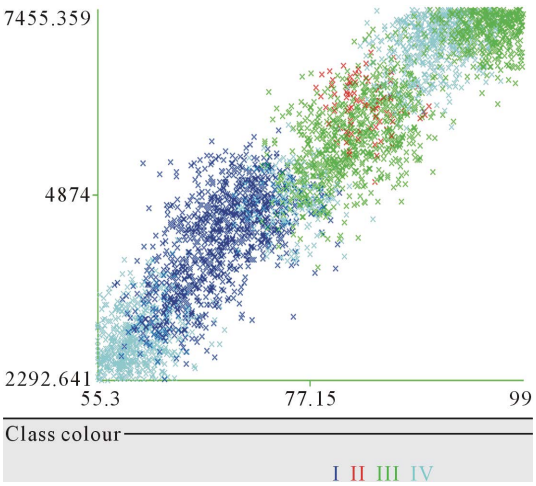

Figure 9. GNP vs time. Class colour correspond to CCA quadrant.

For example, the blue group has a break that positively correlated with GNP (possible correlated to oil boom), then the 80 has another break now correlates negatively with GNP (The 1982 crisis). Change from 80 to 90 now the correlation with GNP is positive (the creation of the National Water Comission—CONAGUA) and mid-90s switch to a negative correlation (the 1994 crisis).

Weka's clustering analysis (**Figures 10-15**) shows that all socioeconomic factors are arranged in two temporal groups: before and after 1989. As for the cluster analysis, due to the high correlation between socioeconomic

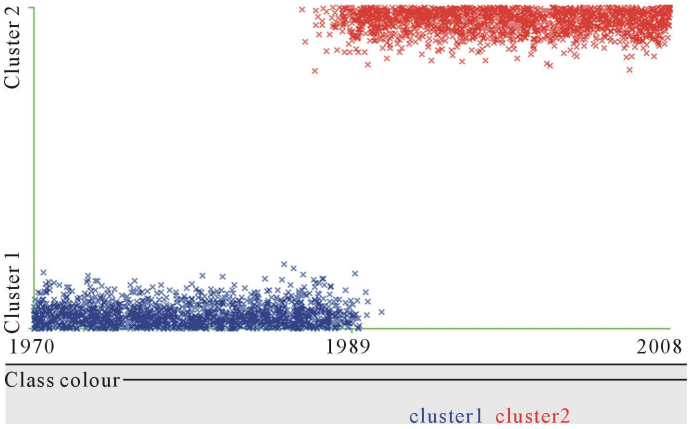

Figure 10. Cluster analysis for years instance.

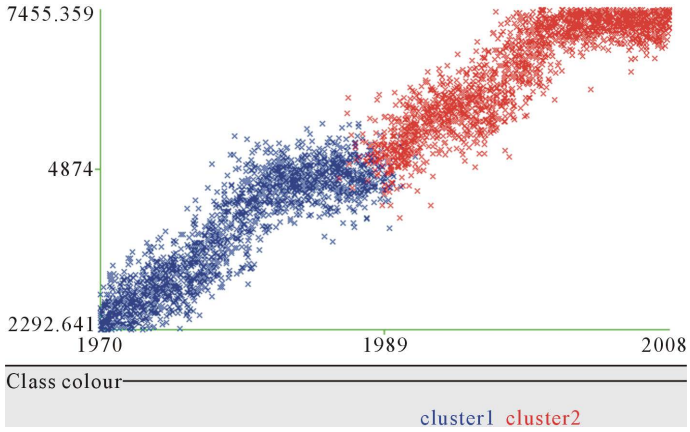

Figure 11. Cluster analysis for GNP.

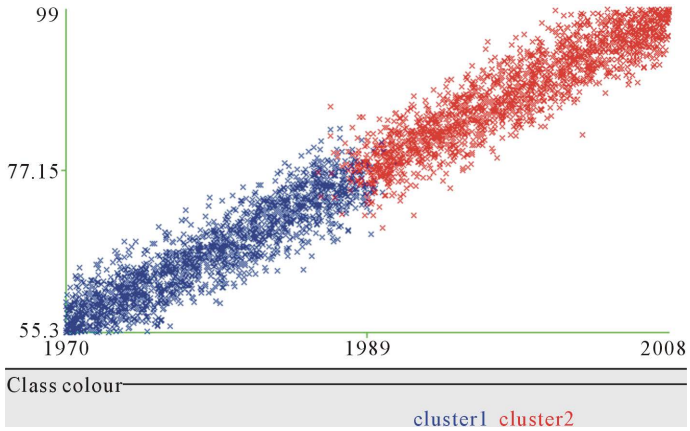

Figure 12. Cluster analysis for state urban population percentage.

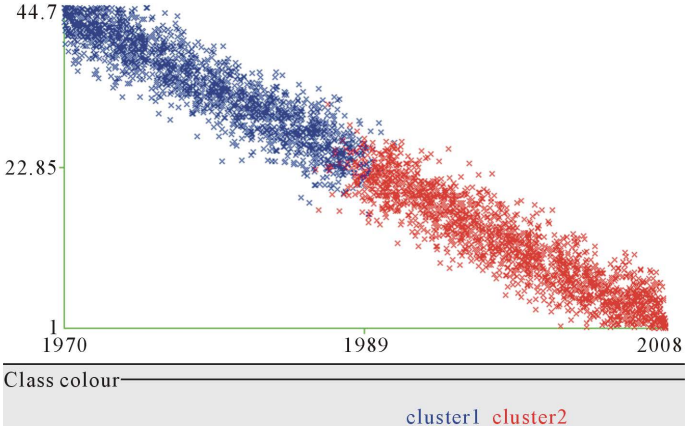

Figure 13. Cluster analysis for state rural population percentage.

```
=== Run information ===

Scheme:       weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S
10
Relation:     mvps
Instances:    3522
Attributes:   13
              año
              s1
              s2
              s3
              s4
              s5
              s6
              s7
              s8
              pob_rur_edo
              pob_urb_edo
              pib
              mvps
Test mode:    evaluate on training data

=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 4
Within cluster sum of squared errors: 2065.6785381157288
Missing values globally replaced with mean/mode

Cluster centroids:
                        Cluster#
Attribute     Full Data         0          1
                 (3522)    (1754)    (1768)
=====================================
año           1988.7073 1979.0319  1998.306
s1              29.6511   28.2184   31.0724
s2              24.4597   22.6261   26.2787
s3              23.7491    22.223   25.2631
s4              20.2587   19.2149   21.2942
s5              19.5622   18.6393   20.4778
s6              15.0216   14.5844   15.4553
s7              10.7046   10.8531   10.5572
s8               7.2578    7.0623    7.4518
pob_rur_edo      23.211    34.337   12.1731
pob_urb_edo      76.789    65.663   87.8269
pib           5201.6905 3821.4465  6571.005
mvps                IV         I        III
```

Figure 14. Weka statistical results for clustering analysis using K-means.

factors of the population and GNP, only this indicator and time were necessary for constructing a decision tree with the algorithm J48 implemented in Weka software by [28] [29].

5. Discussion and Conclusions

Data mining of historic hydrogeological data for the Toluca Valley has proved to be able to generate new knowledge, making clear that groundwater management has been influenced by socioeconomic factors such as GNP and population structure.

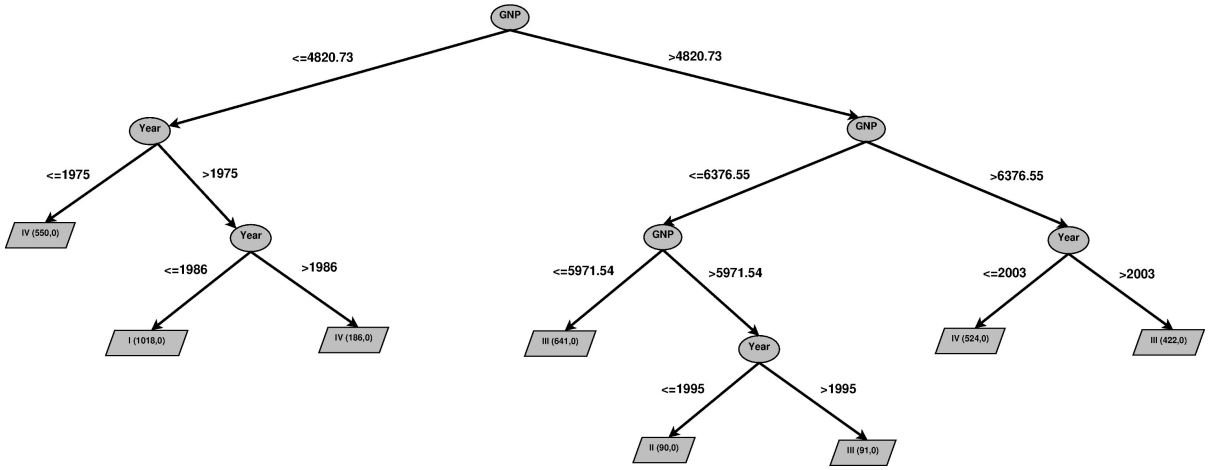

Figure 15. Decision tree for socioeconomic and hydrogoelogical data, using J48 algorithm in Weka.

Interestingly the years used for the algorithm as decision nodes are, with the exception of 2003, years of economic crisis and presidential transitions.

- In 1976, Mexico suffered the consequences of the oil embargo imposed by the Organization of Petroleum Exporting Countries (OPEC) against all the countries that supported Israel in the Yom Kippur War against Syria and Egypt.
- Monday, October 19, 1987, better known as Black Monday, was the day when stock markets around the world crashed, collapsing in a very short time. The crash began in Hong Kong, spread west through international time zones to Europe, hitting the United States after other markets had already declined by a significant margin. Mexico was no exception and this crash devalued the Mexican currency 400% against the US dollar. One year later, in this highly complex economic scenario and with an electoral process that involved two suspicious shutdowns of the computer system used to keep track of the number of votes, makes his entrance Carlos Salinas as President of Mexico.
- The economic crisis that began in Mexico in 1994 had a global impact and was called the "Tequila Effect". It was caused by the lack of international reserves, causing the devaluation of the peso during the first days of the presidency of Ernesto Zedillo, one of the legacies of the Salinas administration. A few weeks before the beginning of the Mexican currency devaluation process, then President of the United States, Bill Clinton, asked the U.S. Congress authorize a credit line for 20 billion US dollars to the Mexican Government.

Even more, Goicochea [30] shows that 1987 and 1989, years that data mining analysis uses for clustering and as one of the year nodes in the decision tree, marks in fact the beginning of the collapse of agropecuarian GNP. These results make it clear that groundwater and economy may be much more linked than thought previously.

We found that hydrogoelogical data evolve in the direction of population transformation from rural to urban. This finding that at first sight may appear as a triviality, may pose a fundamental question about groundwater management. In hydrogeology there is a widely spread thought that groundwater management problems have to do mainly with agricultural issues, such as crop election and irrigation methods. This kind of thinking could mislead us because most of those problems are solvable by agriculture technification. But what happens if, as data mining suggests, groundwater today has much more to do with urban environment? Problems related to cities, such as drinkable water demands, sanitation, distribution, are more complex and possible solutions are more expensive than those required in agriculture. This could represent a whole paradigm shift in groundwater management with profound repercussion in policy making.

Finally it is in general difficult to provide, given a certain regional economic environment, clear and simple criteria that are useful in the design and implementation of suitable private and public policies. Most of the times decisions are taken based on macroeconomic variables and regional and local trends are ignored. Moreover, even when the appropriate relevant variables, have been identified, it is not clear in many occasions how to interpret or establish the causal and interdependence structure among them. Despite the fact that many methodologies are available, it is still true that frequently the results provided by such methods remain difficult to interpret. Data mining as presented in this paper is a useful alternative that gives insights into the dynamics of the system

being studied and helps validating hypothesis that can be helpful in the actual process of decision taking. In this paper, for instance, it is clearly shown that the relationship between rural and urban economic environments is changing in terms of water needs and use and that an appropriate allocation of this resource has to take this change into account. Moreover the connection with the global macroeconomic variables is made apparent in such a way that seasonal changes (in these cases, related to the political situation, namely presidential elections) can be taken into account.

In summary, data mining techniques can provide in relevant economic context a useful methodological alternative by giving simple criteria in decision and policy making.

Acknowledgements

OLC thanks Fondo Capital Semilla at Universidad Iberoamericana and to SNI program with number 62929.

References

[1] Ikoma, E., Taniguchi, K., Koike, T. and Kitsuregawa, M. (2006) Development of a Data Mining Application for Huge Scale Earth Environmental Data Archives. *International Journal of Computational Science and Engineering*, **2**, 262-270. http://dx.doi.org/10.1504/IJCSE.2006.014765

[2] Cios, K.J., Pedrycz, W. and Swiniarski, R.W. (1998) Data Mining and Knowledge Discovery. Springer US, 1-26.

[3] Hand, D. (1998) Data Mining: Statistics and More? *The American Statistician*, **52**, 112-118.

[4] Kantardzic, M. (2003) Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, Hoboken.

[5] Baldi, P. and Brunak, S. (1998) Bioinformatics—The Machine Learning Approach. MIT Press, Cambridge, MA.

[6] Anandhavalli, M., Ghose, M. and Gauthaman, K. (2010) Mining Spatial Gene Expression Data Using Negative Association Rules. arXiv:1001.1991v1 [cs.DB]

[7] Birkholtz, L., Bastien, O., Wells, G., Grando, D., Joubert, F., Kasam, V., Zimmermann, M., Ortet, P., Jacq, N., Roy, S., Hoffmann-Apitius, M., Breton, V., Louw, A. and Maréchal, E. (2006) Integration and Mining of Malaria Molecular, Functional and Pharmacological Data: How Far Are We from a Chemogenomic Knowledge Space? arXiv:q-bio/0611053v1 [q-bio.QM]

[8] Shan, Y. (2006) Genome-Wide EST Data Mining Approaches to Resolving Incongruence of Molecular Phylogenies. arXiv:q-bio/0609004v2 [q-bio.GN]

[9] Fayyad, U., Djorgovski, S. and Weir, N. (1996) Automating the Analysis and Cataloging of Sky Surveys. In: Fayyad, U.M., *et al.*, Eds., *Advances in Knowledge Discovery and Data Mining*, AAIT Press and MIT Press, 471.

[10] Ball, N. and Brunner, R. (2009) Data Mining and Machine Learning in Astronomy. arXiv:0906.2173v1 [astro-ph.IM]

[11] Borne, K. (2009) Scientific Data Mining in Astronomy. arXiv:0911.0505v1 [astro-ph.IM].

[12] Vaduvescu, O., Curelaru, L., Birlan, M., Bocsa, G., Serbanescu, L., Tudorica, A. and Berthier, J. (2009) EURONEAR: Data Mining of Asteroids and Near Earth Asteroids. *Astronomische Nachrichten*, **330**, 698-707. http://dx.doi.org/10.1002/asna.200811198

[13] Karimabadi, H., Sipes, T., White, H., Marinucci, M., Dmitriev, A., Chao, J., Driscoll, J. and Balac, N. (2007) Data Mining in Space Physics: MineTool Algorithm. *Journal of Geophysical Research*, **112**, A11215. http://dx.doi.org/10.1029/2006ja012136

[14] Lavraã, N., Keravnou, E. and Zupan, E. (1997) Intelligent Data Analysis in Medicine and Pharmacology. Kluwer, Alphen aan den Rijn.

[15] Kormushev, P. (2009) Visual Approach for Data Mining on Medical Information Databases Using Fastmap Algorithm. arXiv:0904.0313v1 [cs.IR].

[16] Yang, Y., Cai, X. and Herricks, E. (2008) Identification of Hydrologic Indicators Related to Fish Diversity and Abundance: A Data Mining Approach for Fish Community Analysis. *Water Resources Research*, **44**, W04412. http://dx.doi.org/10.1029/2006wr005764

[17] Bui, E., Henderson, B. and Viergever, K. (2009) Using Knowledge Discovery with Data Mining from the Australian Soil Resource Information System Database to Inform Soil Carbon Mapping in Australia. *Global Biogeochemical Cycles*, **23**, GB4033. http://dx.doi.org/10.1029/2009gb003506

[18] Dhanya, C. and Nagesh, D. (2009) Data Mining for Evolution of Association Rules for Droughts and Floods in India Using Climate Inputs. *Journal of Geophysical Research*, **114**, D02102. http://dx.doi.org/10.1029/2008jd010485

[19] Ailamaki, A., Faloutsos, C., Fischbeck, P., Small, M. and Van Briesen, J. (2003) An Environmental Sensor Network to Determine Drinking Water Quality and Security. *ACM SIGMOD Record*, **32**, 47-52.

http://dx.doi.org/10.1145/959060.959069

[20] Ekasingh, B., Ngamsomsuke, K., Letcher, R. and Spate, J. (2005) A Data Mining Approach to Simulating Farmers' Crop Choices for Integrated Water Resources Management. *Journal of Environmental Management*, **77**, 315-325. http://dx.doi.org/10.1016/j.jenvman.2005.06.015

[21] Scaringella, A. (1999) A Data Mining Application for Monitoring Environmental Risks. In: Perner, P. and Petrou, M., Eds., *Machine Learning and Data Mining in Pattern Recognition*, *MLDM*'99, Lecture Notes in Computer Science, Vol. 1715, Springer, Berlin, 209-215. http://dx.doi.org/10.1007/3-540-48097-8_17

[22] INEGI (1994) Estadísticas de Toluca. Cuaderno Estadístico Municipal. Estado de Mexico. pp. 1, 9.

[23] INEGI (1996) Conteo 1995 Estados Unidos Mexicanos, resultados preliminares.

[24] EDOMEX (2008) Enciclopedia de los Municipios de Mexico Estado de Mexico Toluca de Lerdo. http://inafed.gob.mx/work/enciclopedia/EMM15mexico/index.html

[25] Comisión Nacional del Agua Subdirección General Técnica Gerencia de Aguas Subterráneas Subgerencia de Evaluación y Modelación Hidrogeológica. Determinación de la disponobilidad de agua en el acuífero Valle de Toluca, Estado de México. CNA, 2002.

[26] Fayyad, U., Piatetsky-Shapiro, G. and Smith, P. (1996) From Data Mining to Knowledge Discovery in Databases. *AI Magazine*, **17**, 37-54.

[27] Fellegi, I. and Sunter, A. (1969) A Theory for Record Linkage. *Journal of the American Statistical Association*, **64**, 1183-1210. http://dx.doi.org/10.1080/01621459.1969.10501049

[28] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. (2009) The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, **11**, 10-18. http://dx.doi.org/10.1145/1656274.1656278

[29] Witten, H. and Frank, E. (2000) Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann Publishers, Burlington.

[30] Goicoechea, J. (1996) Modernización y estancamiento: Paradojas del sector agropecuario en México. Comercio Exterior.