Spatial Data Mining of a Population-Based Data Warehouse of Cancer in Mexico

Joaquín Pérez O. [1], M. Fátima C. Henriques[2], Rodolfo Pázos R. [1], Laura Cruz R. [3], Jesús Salinas C.[1], Adriana Mexicano S.[1]
[1]Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), México A.P. xxx– Cuernavaca, México
[2]Secretaría de Saúde do Estado de Pernambuco, Brasil
[3]Instituto Tecnológico de Ciudad Madero, Mexico
{jperez, pazos, jsalinas_05c, iscmexs05c}@cenidet.edu.mx, fhenriques@saude.pe.gov.br, lcruz@itcm.edu.mx

Abstract

This work addresses the problem of discovering patterns of interest in population databases for cancer. In particular, the experimental results obtained by a data mining system are shown, which was developed specifically for this type of databases. The k-means algorithm was used for the generation of patterns, which permits expressing patterns as regions or groups of districts with affinity in their localization and mortality rate parameters. The source databases used in this investigation were obtained from Mexican official institutions. As a result of the application of the system, a set of grouping patterns was generated, which defines the mortality distribution for stomach cancer in Mexican districts.

Keywords: Spatial Data Mining, clustering, Population-based data, cancer.

1. Introduction

The purpose of data mining is to obtain unknown patterns from massive or large databases, which can potentially have a large value or interest for an organization [1], [2].

The diagnosis and treatment of cancer, on one hand, is expensive and absorbs an important part of the public health budgets of many countries including Mexico, and on the other hand, it has a large social impact since the patient's and his relative's lifestyles are affected due to the health care required by the patient. To know the standards of mortality and the space distribution of the illness between different regions and districts of the country, contributes to identify hypotheses to probable causal associations. It still favors to direct measure of prevention and control based in the regional differences identified, and to reduce diagnosis and treatment costs. The standards identified for the present study can direct futures epidemiological studies about stomach cancer mortality. The present method could be extended for others cancers and illnesses.

The occurrence of the malignant tumors of stomach is strongly related to the social economic standards of the populations [3], [4], [5]. This work shows the results obtained by applying a data mining system to a real database of stomach cancer mortality from Mexico. The system was developed ad hoc and consists of a pattern generator subsystem and a visualization subsystem.

1.1 Work related to the application of data mining techniques to health-related databases

In recent years the use of data mining applied to cancer clinical data has increased, some examples are the works in [2], [6], [7], [8], [9]. However, the application of data mining to cancer epidemiologic data has been very limited [10].

In [10] a study is reported on the application of data mining to the analysis of epidemiologic data, where, specifically, the following techniques are mentioned: Classification and Regression Trees, Multivariate Adaptive Regression Splines, and Tree-Structured Classifiers. That paper presents interesting references on the application of data mining techniques. The paper concludes that the application of data mining techniques to population databases has been limited and that its use may facilitate the finding of interesting patterns for this kind of data.

According to the specialized literature surveyed, no previous works have been reported where grouping techniques are applied to population-based data on cancer for obtaining mortality rate distributions.

2. Data Mining Methodology

2.1 Source of Data

In this research, real data from several official databases were used. The most important are described hereupon.

a) Mortality data for cancer

This data was extracted from the *Núcleo de Acopio y Análisis de Información de Salud NAAIS* (Collection and Analysis Core on Health Information) [11] from the *Instituto Nacional de Salud Pública INSP* (National Institute for Public Health). From this database, all the records on deaths from stomach cancer for the year 2000 were selected, and out of 38 attributes of the table only two were included: death cause and deceased district.

b) Population and geographic data

The data on districts population were obtained from the *Sistema Municipal de Bases de Datos SIMBAD* (Database District System) [12] from INEGI. The data on the geographical position of each district and maps of Mexico were also obtained from INEGI.

2.2 Data preprocessing

For each district, the gross stomach cancer mortality rate per 100,000 inhabitants for the year 2000 was calculated, using formula (1).

$${}_{00}\,rate = \frac{death}{population} * 100{,}000 . \qquad (1)$$

where:

death : number of stomach cancer death at a district in the year of 2000.

population: district population for the year 2000.

The data on geographical position of districts was transformed, specifically minutes and seconds, to their equivalent in fractions of degree.

As a result of the preprocessing, the data warehouse was populated for the application of several modeling techniques [13].

2.3 Data modeling techniques

For knowledge extraction, the data was modeled through grouping. To this end the K-means algorithm was chosen because its results were satisfactory for the problem under consideration. Patterns were generated as groups of districts with similar parameters regarding geographical position and mortality rate. Weka [14] was used for the experimentation.

2.4 Generation of geographical group patterns

The results of the grouping algorithm are lists of element groups including the centroid of each group. Since the interpretation of patterns expressed in list form was difficult for specialists, it was considered convenient to present groups information in tabular form

(Tables 1, 2, 3); however, this was not entirely satisfactory, since there exist over 1500 districts in Mexico and it is not easy to locate a district on a map, specially if it is relatively unknown. In order to solve this problem with interpretation, a visualization subsystem was developed, which is described in the following section.

2.5 Pattern visualization

The cartographic visualization subsystem permits selecting and drawing on a map of Mexico one or more of the patterns generated by the grouping algorithm. The subsystem shows on the map the group centroids as small circles and the group elements (districts) as black dots; while the membership of an element to a group is indicated by a line that joins it to the group centroid (Figs. 1,.., 5).

The visual representation of groups permitted to enhance the knowledge obtained and facilitated the interpretation and assessment of the results by the system users.

3. Experimental results

A set of experiments were conducted using the data mining system on the cancer data warehouse, selecting districts with population greater than 100,000 for the year 2000, and setting the number of groups *k* equal to 5, 10, 15, 20, and 30. The best result was obtained for *k* equal to 20 according to the specialists. Figure 1 shows a map of Mexico depicting the political division by states and districts. The figure shows the set of 20 group patterns, which defines the distribution of the gross mortality rates for stomach cancer per 100,000 inhabitants for Mexican districts.

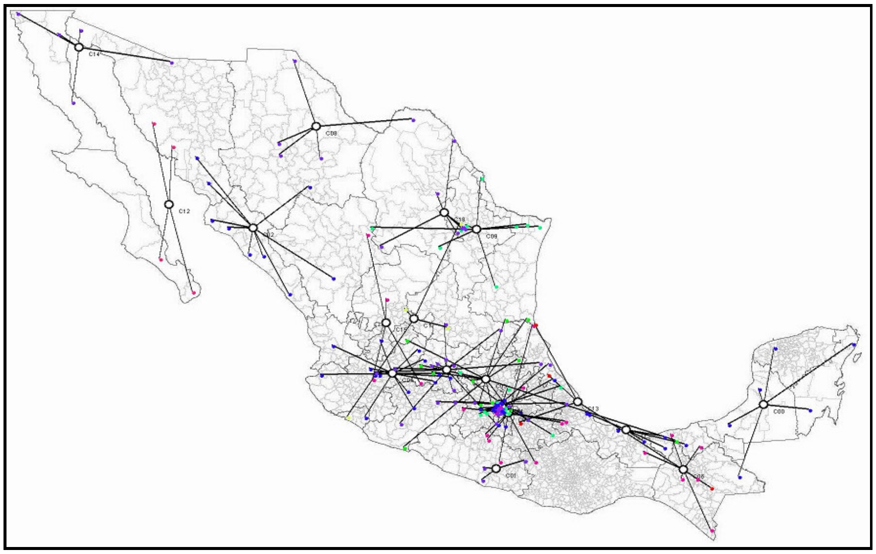

Figure 1. Mortality groups for stomach cancer (C16) for year 2000

Out of the 20 groups generated, the three groups with the largest average rates were selected for attracting the most interest. Figure 2 shows group three on the bottom right corner, which corresponds to the Chiapas heights in the southeast of Mexico. This group served to validate the data mining method used in this prototype, since clinical investigations have reported a high mortality rate for gastric cancer. Such investigations have claimed that one of the factors that contribute to the development of this type of cancer in the region is a chronic infection caused by a bacteria called helicobacter pylori (HP) [15]. The district details of the group and the mortality rates are shown in Table 3 including the mean value and the standard deviation. Figure 5 shows a close up of the southeast of Mexico including the graphic representation of group three.

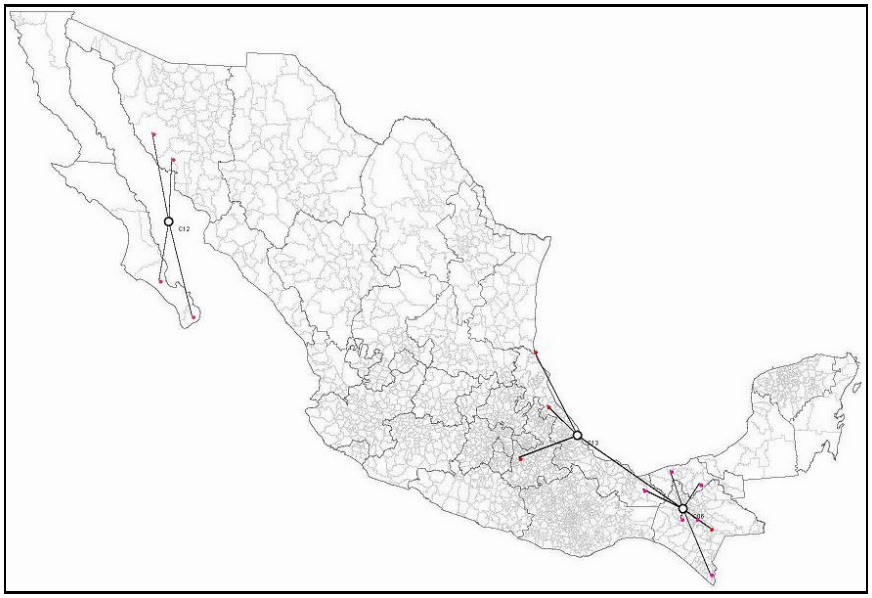

Figure 2. Three interesting patterns found for cause C16 for year 2000

Additionally, we report as a new finding another pattern of interest and potential usefulness in the northwest region: group two (Fig. 4, Table 2), which has an average mortality rate larger than that of group three. According to the specialized literature, there are no studies reporting a concentration of high mortality rates for stomach cancer in this region. A possible explanation for this situation is that cancer statistics are usually analyzed statewise and group two spans two states.

Group one (Fig. 3, Table 1) is the one with the largest average mortality rate; however, we do not consider it meaningful, since the location of its member districts is more sparse than that of the other groups.

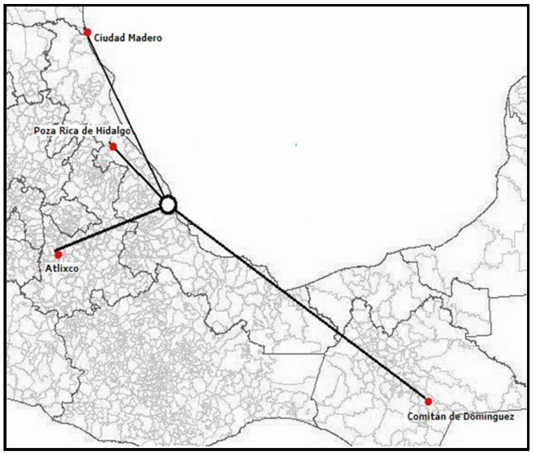

Figure 3. Group 1

Table 1. Group 1

District	Rate
Comitán de Domínguez	11.41
Poza Rica de Hidalgo	11.12
Atlixco	10.25
Ciudad Madero	9.87

Average	10.66
Standard deviation	0.62

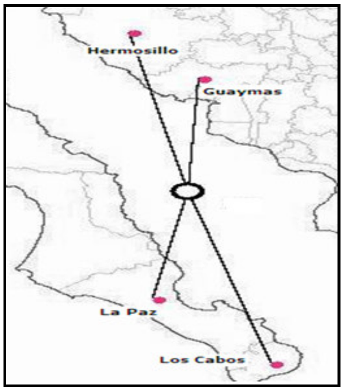

Figure 4. Group 2

Table 2. Group 2

District	Rate
Guaymas	11.52
Hermosillo	7.87
La Paz	7.11
Los Cabos	6.64

Average	8.28
Standard deviation	1.92

Table 3. Group 3

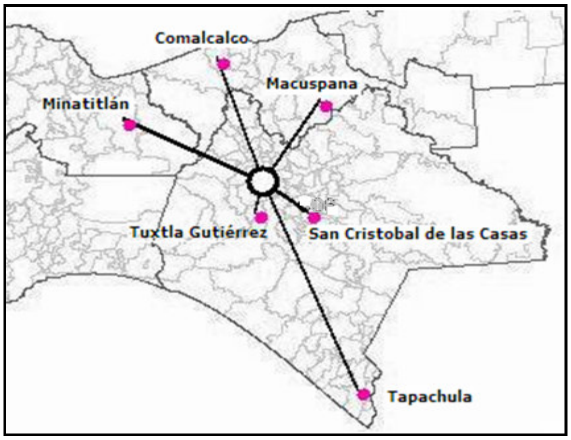

Figure 5. Group 3

District	Rate
Minatitlán	9.15
Comalcalco	8.50
Tapachula	7.73
San Cristóbal	6.80
Macuspana	6.72
Tuxtla Gutiérrez	6.45

Average	7.56
Standard deviation	0.99

4. Conclusions

One of the most important contributions of this work is the development of a population-based data warehouse on cancer (stomach, lung, etc.) from official data sources, and specifically the integration of geographical data of districts with cancer statistical data.

K-means was used as a grouping algorithm, which proved to be an adequate option for grouping geographical regions. A geographic visualization subsystem was implemented, which permitted to show the centroid and the districts of groups on a map, allowing to depict patterns as nation regions with similar mortality rates. This tool proved to be particularly useful for assessing and communicating the results because its visual expressiveness.

A set of experiments were conducted using the data mining system for different numbers of groups (*k*), and *k* = 20 yielded the best result. Figure 1 shows the set of 20 patterns that defines the distribution of the gross mortality rates per 100,000 inhabitants for stomach cancer in the districts of Mexico for the year 2000.

As a result of the analysis of the patterns generated for stomach cancer, a well known pattern of districts with high mortality rate in the southeast of Mexico was determined (Fig. 5, Table 3), which served for validating the method used in this prototype.

Additionally, we report as a new finding another pattern of interest and potential usefulness in the northwest region: group two (Fig. 4, Table 2), which has an average mortality rate larger than that of group three. Group one (Fig. 3, Table 1) has the largest average mortality rate; however, the localization of its districts is more sparse than that of groups two and three, and therefore we do not consider group one meaningful.

We consider that our data mining system can be improved by developing functions for adjusting mortality rates by age intervals and gender. Another improvement could consist of the integration of modules for the analysis of mortality rates for other diseases besides cancer.

Finally, we consider that the patterns generated by the data mining system, which are expressed as groups of districts with similar location and mortality rate parameters, can be useful as an aid tool for studies on cancer and for decision making concerning the allocation of resources for organizing specialized services for cancer prevention and treatment.

References

1. Larose T. D. (2006), Data mining methods and models, John Wiley & Sons, New Jersey.
2. Thangavel K., Jaganathan P., and Esmy P. (2006),“Subgroup Discovery in Cervical Cancer Analysis using Data Mining”, AIML Journal, Volume (6), Issue (1), January, p. 29-36.
3. Faggiano F, Partanen T, Kogevinas M, Boffetta P. (1997), “Socioeconomic differences in cancer incidence and mortality”, In: Social inequalities and cancer, Edited by Kogevinas M, Pearce N, Susser M, Boffetta P, editors. Social inequalities and cancer, (138):65-176. International Agency for Research on Cancer (IARC), Lyon, France.
4. Bouchardy C, Parkin DM, Khlat M, et al. (1993), “Education and mortality from cancer in São Paulo, Brazil”, Annals of Epidemiology, 3(1):64-70 International Agency for Research on Cancer (IARC), Lyon, France.

5. Nomura A. (1996),"Stomach cancer" In: Cancer epidemiology and prevention , Edited by Schottenfeld D, Fraumeni JF Jr., Oxford University Press, New York.
6. Nevine M. Labib and Michael N. Malek (2005), "Data Mining for Cancer Management in Egypt Case Study: Childhood Acute Lymphoblastic Leukemia", Transactions on Engineering, Computing and Technology V8, p. 309-314.
7. Mullins IM, Siadaty MS, Lyman J, Scully K, Garrett CT, Miller WG, Muller R, Robson B, Apte C, Weiss S, Rigoutsos I, Platt D, Cohen S, Knaus WA. (2006), "Data Mining and clinical data repositories: Insights from a 667,000 patient data set" Computers in Biology and Medicine, p. 1351-1377.
8. Wheeler D. (2007), "A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996-2003", International Journal of Health Geographics, 6:13.
9. Maheswaran R., Strachan D., Dodgeon B. and Best N. (2002), "A population-based case-control study for examining early life influences on geographical variation in adult mortality in England and Wales using stomach cancer and stroke as examples", International Journal of epidemiology, p. 375-382.
10. Flouris A. & Duffy J. (2006), "Application of artificial intelligence systems in the analysis of epidemiological data", European Journal of Epidemiology, p. 167-170.
11. NAIIS, Instituto Nacional de Salud Pública, Núcleo de Acopio y Análisis de información en Salud (2003), http://sigsalud .insp.mx/naais/.
12. SIMBAD, Instituto Nacional de Estadística, Geografía e Informática, Sistema Municipal de Base de Datos (SIMBAD) (2007), http://sc.inegi.gob.mx/simbad/index.jsp?c=125.
13. Hernández Orallo José, Ramírez Quintana Ma. José, Ferri Ramírez César (2004) "Introducción a la Minería de Datos", Pearson Educación, Madrid.
14. Ian H. Witten and Eibe Frank (2005) "Data Mining: Practical machine learning tools and techniques", 2nd Edition, Morgan Kaufmann, San Francisco.
15. Mohar A., Ley C., Guarner J., Herrera-Goepfert R., Sánchez L., Halperin D., Parsonnet J. (2002), "Alta frecuencia de lesiones precursoras de cáncer gástrico asociadas a Helicobacter Pylori y respuesta al tratamiento, en Chiapas, México", Gaceta Médica de México, p. 405-410.