Data Warehouse Development to Identify Regions with High Rates of Cancer Incidence in México through a Spatial Data Mining Clustering Task.

Joaquin Pérez Ortega[1],
María del Rocío Boone Rojas[1,2],
María Josefa Somodevilla García[2],
Mariam Viridiana Meléndez Hernández[2]

1 Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca Mor. Mex.
2 Benemèrita Universidad Autónoma Puebla, Fac. Cs. de la Computaciòn, México.
jperez@cenidet.edu.mx,{rboone,mariasg}@cs.buap.mx,*mvmh_099@hotmail.com*

Abstract

Data warehouses arise in many contexts, such as business, medicine and science, in which the availability of a repository of heterogeneous data sources, integrated and organized under a unified framework facilitates analysis and supports the decision making process. These data repositories increase their scope and application, when used for data mining tasks, which can extract useful knowledge, new and valuable from large amounts of data.

This paper presents the design and implementation of population-based data warehouses on the incidence of cancer in Mexico; based on the conceptual level multidimensional model and the ROLAP model *(Relational On-Line Analytical Processing)* at the implementation level.

A data warehouses is built, to be used as input for clustering data mining tasks, in particular, the k-means algorithm, in order to identify regions in Mexico, with high rates of cancer incidence.

The identified regions, as well as, the dimension related to the geographic location of the municipalities and their rate of incidence of cancer, are processed by IRIS, a Geographic Information System, developed at the National Institute of Statistics, Geography and Informatics of Mexico.

1 Introduction

Data warehouses arise in many contexts, such as business, medicine and science, in which the availability of a repository of heterogeneous data sources, integrated and organized under a unified framework facilitates analysis and supports the decision making process. These data repositories increase their scope and application, when used for data mining tasks, which can extract useful knowledge, new and valuable from large amounts of data.

Data warehouses have been applied mainly in the commercial and business areas [3] and more recently there have been some applications in the Health field

[16] [17] and the trend towards its integration with various technologies [11] [16].

Moreover, according to the literature, the use of data mining systems applied to the analysis of massive databases of health on a population basis has been limited, it is noteworthy work: *Constructing Over Dendrogram Matrix Detail view + Views.* [6], *Application of data mining techniques to databases population of cancer* [1], *Subgroup discovery in cervical cancer using data mining Techniques* [18] and *Data mining for cancer management in Egypt* [10]. In the case of Mexico, to the best of our knowledge, the work that has been developed at the *Centro Nacional de Investigación y Desarrollo Tecnológico* and BUAP, are the first ones in this field.

This work has been preceded by other works which has been done on the incidence of other cancers such as stomach and lung [15]. It is part of a larger project doomed to make proposals for improving the k-means algorithm in various aspects such as effectiveness and efficiency, reported in [12], [13] and [14] and its application in the Health field.

This article presents the data warehouse design and integration for the development of a data mining task on cancer incidence by regions in Mexico, based on the integration of complementary technologies such as clustering and geographical information systems. As a study case, the results for the incidence of cervical cancer are presented, which has been of special interest, since in Mexico, cervical cancer is the leading cause of cancer death in women [11].

The report is organized as follows, followed by this introduction, Section 2 presents the description of data sources and process design and implementation of data warehouse, Section 3 provides an overview of each application. In Section 4, results for the case of cervical cancer and its visualization by GIS INEGI IRIS [5] are included. Finally, in Section 5, conclusions and perspectives of this work are presented.

2 The Data Warehouse

The process of collecting and integrating data warehouse on cancer incidence by region in Mexico, required to select the data sources necessary to accomplish the task of data mining. This section describes the data sources and the conceptual design based on the multidimensional model and also, the implementation of the data warehouse under the ROLAP approach.

2.1 The Data Sources

In the study, the processed databases have been derived from official records of the National Institute of Public Health (INSP) and the National Institute of Statistics, Geography and Informatics (INEGI) of Mexico.

Data on cancer incidence were obtained through subsystem Remote Consultation System for Health Information (SCRIS) of the INSP [9]. In particular, the databases were queried for cases of mortality cancer and results were configured by considering levels of aggregation such as: National States, division (Jurisdiction, Municipalities), year, age range, gender and causes (including tumors).

The information on the population and the actual geographical location of the municipalities was obtained from INEGI official databases, through its Geographic Information System IRIS, which has statistical information covering a wide geographical number of subjects, demographic, social and economic; also includes aspects of the physical environment, natural resources and infrastructure. This wealth of statistical and geographical data was obtained through various activities such as conducting population and housing census and economic census and the generation of basic cartography and census.

The information in the databases of the above institutions are integrated into a data warehouse (see Fig. 1), and according to the conventions in the area of health, for this study, only the municipalities with more than one hundred thousand inhabitants were considered.

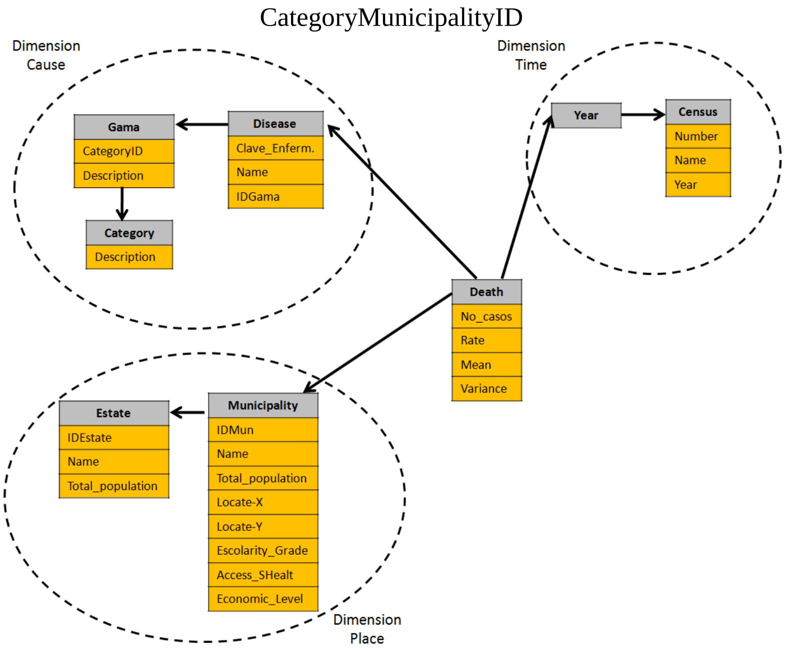

Fig. 1 Multidimensional Model Data Warehouse on the incidence of cancer in Mexico.

2.2 Data Warehouse Multidimensional Model for a population-based incidence of cancer in Mexico.

According to [4] the conceptual data model most widely used for data warehouses is the multidimensional model. The data are organized around the *facts* that have attributes or *measures* that may be more or less detail according to certain *dimensions.* In our case, the data warehouse design at the conceptual level is based on the multidimensional model, in which the dimensions can be distinguished as CAUSE, TIME, and PLACE. In this case, it is considered that a country has the basic fact, "deaths" that may have associated attributes such as number of cases, incidence rate, mean, variance, etc.. Fact can be detailed in several dimensions such as cause of death, place of death, date of death, etc. In Fig. 1 shows the facts "deaths" and three dimensions with various levels of aggregation. The arrows can be read as "is added". As shown in Fig. 1, each dimension has a hierarchical structure but not necessarily linear. When the number of dimensions cannot exceed three represent each combination of levels of aggregation as a cube.

The cube is made up of boxes with one box for each possible value from each dimension to the corresponding level of aggregation. On this "view", each box represents a fact. Fig. 2 shows a three dimensional cube corresponding to the fact: "According to the 2000 census, the town of Atlixco, there were 15 deaths from cervical cancer" in which the dimensions Cause, Place and Time have been added by type of disease (cancer), Municipality and Census. The representation of a fact corresponds therefore to a square in the cube. The value of the box is the observed (in this case is the number of deaths).

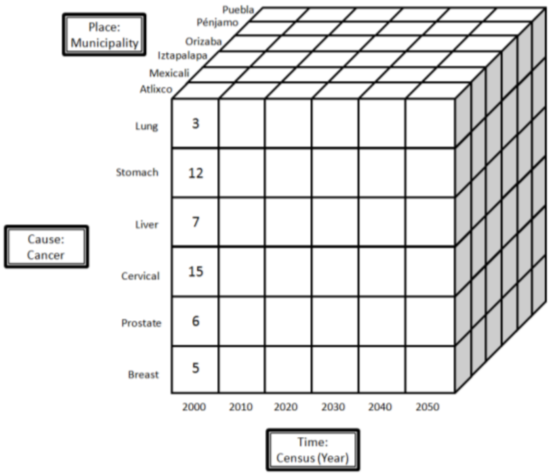

Fig. 2 Display of a fact in a multidimensional model

2.3 Data warehouse scheme ROLAP (Relational OLAP) implementation of population-based cancer incidence in Mexico.

One of the most efficient ways to implement a multidimensional model using relational databases is based on the ROLAP model [4]. In our case, the tables for the ROLAP model have the following schemes:

Snowflake Tables

Dimension Cause

DISEASE (Clave_Enfermedad, name, IdGama, CategoryID)
GAMA (IdGama, CategoryID, Description)
CATEGORY (CategoryID, Description)

Place dimension

STATE (Clave_Estado, name, población_total)
MUNICIPALITY (Clave_Municipio, Clave_Estado, name, población_total, Loc_x, Loc_y, extension, tipo_zona, nivel_socioeconómico)

Time dimension

YEAR (Idan)
CENSUS (IdCenso, Idan, number, name)

Fact Tables

DEATH (IdEnfermedad, IdCenso, IdMunicipio, no_casos, rate, mean, variance)

Star Tables

TIME (Idan, IdCenso)
CAUSE (IdEnfermedad, IdGama, CategoryID)
PLACE (IdCiudad, IdMunicipio)

3 Data Mining Application on Cancer Incidence

The implemented data warehouse has been used to develop a data mining task space based on the integration of additional technologies to the data warehouse, such as clustering and Geographic Information Systems, which in this case are very suitable, to identify and display areas with incidence of cancer in Mexico. The following provides a general description of the integration process of technologies and tools (Fig. 3) made for this application.

The data warehouse integrates the following information for our application: the component space that allows viewing of the regions of municipalities, population data such as the death rate and incidence rate and the time component, which in this case is the census year.

The IRIS GIS INEGI [5], through your options allows the recovery of population data and the real location of the municipalities, which are integrated into the data warehouse.

Since IRIS stores geographical representation of municipalities in the vector format standardized *"shape"* and by means of polygons, there is the need for a process of transfer of forms and formats in order to have a numerical representation of each municipality, in this case, corresponds to a point on the municipality center location, which is accomplished primarily through the tools of ESRI's ArcInfo GIS.

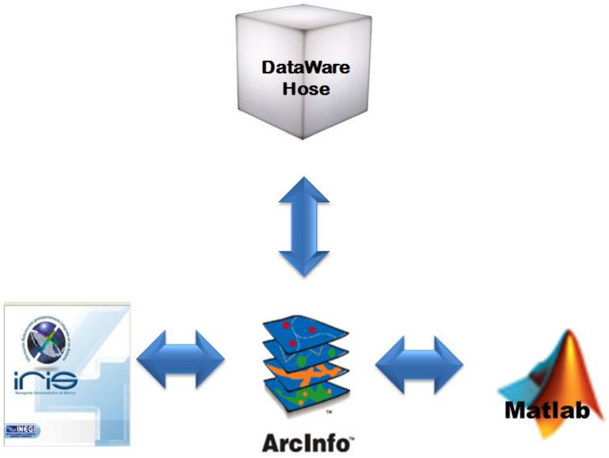

Fig. 3 Integration of Technology and Data Mining Tools

Given the numerical representation of each municipality through a point (x, y), along with its rate of incidence of cancer, the Matlab programming environment and its implementation of *k-means* algorithm [2] [7] is used *to* generate patterns / groups of municipalities and the corresponding centroids.

Once you have the above results, it is again necessary to transfer digital data format to format *shape*, a process similar to above using ArcInfo tools, allowing viewing through GIS IRIS.

Finally, the groups of municipalities and their corresponding centroids, are passed as GIS layers to IRIS, for display on the geographic map of Mexico.

4 Results and visualization with IRIS

In this project we have done grouping tasks according to the affinity of location and incidence rate of the municipalities. Series of experimental tests on the data stores in cities with more than 100.000 inhabitants were carried out. Size groups were considered k = 5, 10, 15, 20 and 30. The best result was obtained for k = 20.

As a case study, this paper presents the results obtained by k-means algorithm in Matlab for the cervical cancer data warehouse. Fig. 4 provides the visualization of the 20 regions identified.

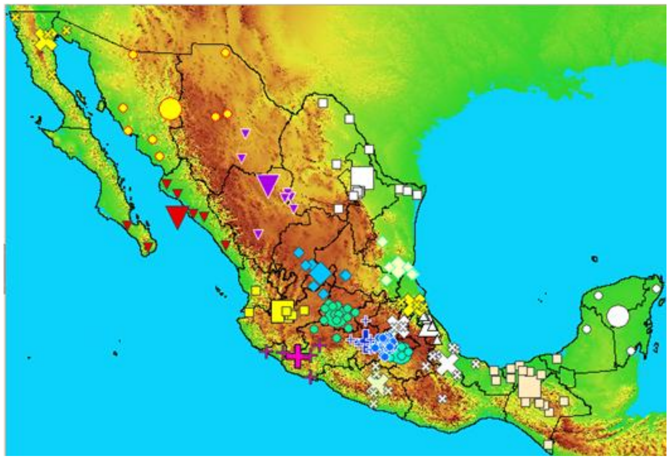

Fig. 4 Regions of the Municipalities with an incidence of Cervical Cancer.

From the results, we distinguish the groups spearheading the three municipalities with higher incidence rates: *Atlixco*, *Apatzingán* and *Tapachula (Chiapas)*. In Fig. 5 the detail of the display of the group corresponding to the region of Chiapas and the incidence of cervical cancer is shown. Table 1 provides data for the previous group, and statistical measures for the mean and standard deviation.

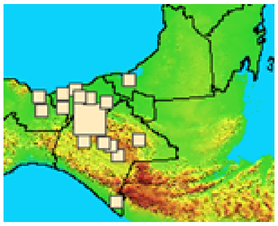

Fig. 5 Tapachula Chiapas Group

The groups identified with high incidence rates: *Tapachula* and *Apatzingan* match municipalities identified in other studies [4] and correspond to the population characteristics, identified in the work of the medical field [8], [15] such situations such as poverty, lack of preparation and access to effective health services and the initiation of sexual activity at an early age. This allows us to assert that the grouping is made valid. On the other hand, the study allowed discovering other municipalities that had not been identified in other research, such as the group of Atlixco, in particular showing the highest incidence rate in the country (see table 2).

Table 1 Municipalities Incidence Rates of Cervical-Uterine Cancer

State	Municipality	Population	Deaths	Rate
Chiapas	Tapachula	271674	27	9.93
Veracruz-Llave	Coatzacoalcos	267212	23	8.60
Veracruz-Llave	Minatitlán	153001	13	8.49
Chiapas	Comitán de Domínguez	105210	8	7.60
Chiapas	San Cristóbal de las Casas	132421	9	6.79
Tabasco	Comalcalco	164637	11	6.68
Tabasco	Cárdenas	217261	11	5.06
Tabasco	Huimanguillo	158573	8	5.04
Chiapas	Tuxtla Gutiérrez	434143	21	4.83
Tabasco	Cunduacán	104360	5	4.79
Campeche	Carmen	172076	8	4.64
Tabasco	Macuspana	133985	6	4.47
Tabasco	Centro	520308	23	4.42
Chiapas	Ocosingo	146696	2	1.36
Average				5.91
Standard deviation				2.23

In order to perform a global analysis of our results, Table 2 provides information of the ten municipalities with the highest incidence rate in the country.

Table 2 Top Ten Municipalities Incidence Rates of Cervical-Uterine Cancer

Key	State	Municipality	Population	Deaths	Rate
21019	Puebla	Atlixco	117111	15	12,80
16006	Michoacán	Apatzingán	117949	13	11,02
07089	Chiapas	Tapachula	271674	27	9,93
17006	Morelos	Cuautla	153329	14	9,13
28021	Tamaulipas	El Mante	112602	10	8,88
06007	Colima	Manzanillo	125143	11	8,78
30039	Veracruz-Llave	Coatzacoalcos	267212	23	8,60
18017	Nayarit	Tepic	305176	26	8,51
30108	Veracruz-Llave	Minatitlán	153001	13	8,49
30118	Veracruz-Llave	Orizaba	118593	10	8,43
General Mean					4.70
Standard Deviation					1.95

Figure 6, illustrates the location of previous incidence rates compared to the national average and the corresponding standard deviation.

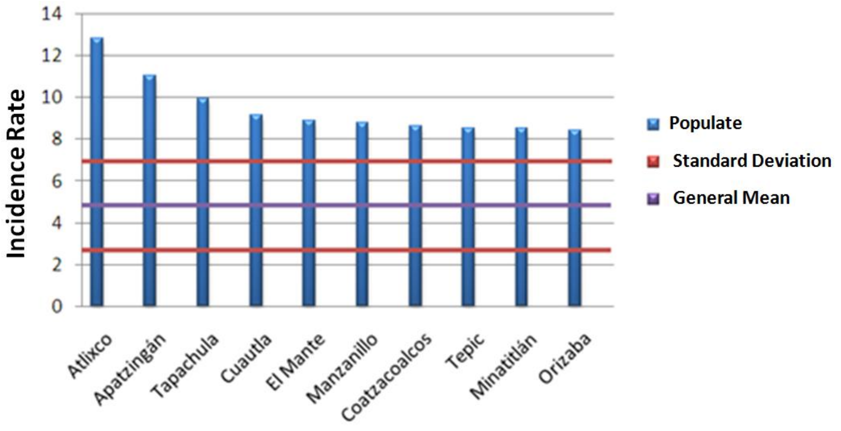

Figure 6. Top Ten municipalities incidence rates.

5 Conclusions

Multidimensional model for conceptual design of the data warehouse, turned out to be very appropriate, since this model is easily scalable and allows analysis of the information under different perspectives. It is expected that future studies process other variables, related to the municipalities, included in this design, such as socioeconomic status, type of region, gender and access to health services, among others. Moreover, the implementation of data warehouse based on the ROLAP model has allowed taking advantage of the facilities developed for relational databases. In addition, it is expected that the design and implementation carried out in the data warehouse can be used in other applications.

The processing of the spatial component of our data warehouse, using the IRIS GIS INEGI, has resulted in a high quality visual representation of our results, based on the actual physical location of the municipalities and on a map of the topography of the Republic Mexican INEGI. Also experience and learning has been gained on transfer of shapes (polygons, points) techniques and formats *(Number-shape)* through ArcView GIS tools.

Currently we are working to complete studies in other cancer types. Besides, data mining tasks will be developed on the incidence of conditions such as diabetes, influenza and cardiovascular diseases, among others.

Acknowledgement. R. Boone expresses her gratitude to Ms. Rocío Pérez Osorno from INEGI, Puebla. (Graduated from the Faculty of Cs. Computing, BUAP) for advice and support in plotting the results of this work through the IRIS GIS.

References

1. Barrón Vivanco M. Arandine, Pérez O. J., Miranda H. Fátima, Pazos R., XII Congreso de Investigación en Salud Pública, *Aplicación de técnicas de minería de datos a bases de datos poblacionales de cáncer*, CENIDET, México, Secretaría de Saúde do Estado de Pernambuco, Brasil, Abril (2007).
2. Forgy E. “Cluster analysis of multivariate data: Efficiency vs. Interpretability of classification”, Biometrics, vol. 21, pp.768-780.1965
3. Hernández-Orallo J., Ramiréz-Quintana M. J., Ferri-Ramiréz C., Introducción a la Minería de Datos, Ed. Pearson Prentice Hall, Madrid (2004).
4. Hidalgo-Martínez Ana C. El cáncer cérvico-uterino su impacto en México. Porqué no funciona el programa nacional de detección oportuna. Revista Biomédica, Centro Nal. De Investigaciones Regionales Dr. Hineyo Noguchi, UADY, 2006, México.
5. IRIS 4. http://mapserver.inegi.gob.mx. SNIEG Sistema Nacional de Información Estadística y Geográfica.
6. Jin Chen, MacEachren, Alan M., Peuquet, Donna. Constructing Overview+Detail Dendogram Matrix Views. IEEE Transactions on Visualization & Computer Graphics., Vol. 15, Issue 6, p889-896, Dec. 2009.
7. MacQueen, J.: Some Methods for Classification and Analysis of Multivariate Observations. In Proceedings Fifth Berkeley Symposium Mathematics Statistics and Probability. Vol. 1. Berkeley, CA (1967) 281-297.
8. Martínez M. Francisco Javier. Epidemiología del cáncer del cuello uterino. Medicina Universitaria 2004, 39-46. Vol. 6, N. 22, UANL, México.
9. NAIIS Instituto Nacional de Salud Pública, SCRIS, Mortalidad, http://sigsalud.insp.mx/naais/, Cuernavaca, Morelos, México, (2003).
10. Nevine M. Labib, Michael N. Malek: Data Mining for Cancer Management in Egypt. Transactions on Engineering, Computing and Technology V8 October 2005: (ISSN 1305-5313).
11. Pérez-C. Nelson, Abril-Frade D.O. Estado Actual de las Tecnologías de Bodegas de Datos Espaciales. Ing. E Investigación. Vol.27, No. 1, Univ. Nal. De Colombia. 2007.
12. Pérez-O. J.,1, R. Pazos R, L. Cruz R.,G. Reyes S. “Improvement the Efficiency and Efficacy of the K-means Clustering Algorithm through a New Convergence Condition”. Computational Science and Its Applications – ICCSA 2007 – International Conference Proceedings. Springer Verlag.
13. Pérez-O. J.2, M.F. Henriques, R. Pazos, L. Cruz, G. Reyes, J. Salinas, A. Mexicano. Mejora al Algoritmo de *K-means* mediante un Nuevo criterio de

convergencia y su aplicación a bases de datos poblacionales de cancer. 2do Taller Latino Iberoamericano de Investigación de Operaciones, Mèxico, 2007.

14. Pérez-O. J.3, Rocío Boone Rojas, María J. Somodevilla García. Research issues on K-means Algorithm: An Experimental Trial Using Matlab., Advances on Semantic Web and New Technologies”. Vol 534. http://ceur-ws.org/.

15. Rangel-Gómez, G. Lazcano-Ponce,E. Palacio-Mejía, Cáncer cervical, una enfermedad de la pobreza: diferencias en la mortalidad por áreas urbanas y rurales en México, http:// www.insp.mx/salud/index.html.

16. Scotch,Matthew, Parmato B. Monaco, V. Evaluation of SOVAT: An OLAP-GIS decision support system for community health assessment data analysis. BMC Medical Informatics & Decisión Making Vol. 8 (1-12). 2008.

17. Simonet, A., Landais, P. Guillon D.A multi-source Information System for end-stage renaldisease. Comptes Residus Biologies, 2002, Vol. 325 I4., p515.

18. Thangavel K. Jaganathan P. and Esmy P. O., *Subgroup Discovery in Cervical Cancer Analysis Using Data Mining Techniques*, Departament of Computer Science, Periyar University: Departament of Computer Science and Applications, Gandhigram Rural Institute-Deemed University, Gandhigram: Radiation Oncologist , Christian Fellowship Community Health Centre, Tamil Nadu, India: AIML journal, Vol(6), Issue(1), January, 2006.