Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 1 (2012) 979–986

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference on Computational Science, ICCS 2010

Geomedica: managing and querying clinical data distributions on geographical database systems

G. Tradigo[a,*], P. Veltri[a], S. Greco[b]

[a] *University Magna Græcia of Catanzaro, Catanzaro 88100, Italy*
[b] *University of Calabria, Arcavacata di Rende 87036, Italy*

Abstract

Geographical databases are a significant and mature tool, useful in many application areas thanks to the spread of new positioning and mapping technologies. Geographical functionalities can be added to existing applications, from land management to water and electricity control systems. The use of geographical information applications greatly improves data interpretation, thus helping users in making better decisions. Further improvements can be obtained by using more sophisticated tools (e.g. On Line Analytical Processing and Data Mining techniques) to highlight interesting and previously unknown relations on spatio-temporal data, which can help in a better understanding of data.

In this paper we report the experience of using GIS technologies to analyze clinical data containing health information about a large population. Clinical data have been geocoded by associating tuples related to some geographical position with the coordinates of a map and then analyzed and queried using both SQL-like languages and a graphical user interface. Several experiments have been performed using data related to an italian district which have been furnished by an association of family doctors and patients. Test queries performed on the available dataset were able to correctly correlate health data about patients with geographical features (e.g. points of interest, boundaries, coastlines vectors) and to visualize diseases geographical distributions on a map.

Keywords:
GIS, SIT, Geographical Information Systems, Spatial Databases, Points Of Interest, Spatial Clinical Data, Epidemiological Databases

1. Introduction and Motivation

A Geographical Information System (GIS) is a collection of spatially referenced data (i.e. data that have locations attached to them) and a set of tools which can be used to work with the data. Nowadays we usually refer to GIS as a program running on a computer, more in general we can refer to a GIS even speaking of a properly organized set of paper files on a desktop or in an archive. We will show how a computer-based GIS system can be used to exploit epidemiological datasets.

*Corresponding author
Email addresses: gtradigo@unicz.it (G. Tradigo), veltri@unicz.it (P. Veltri), greco@unical.it (S. Greco)

1877-0509
doi:10.1016/j.procs.2010.04.108

Epidemiological phenomena, as well as new virus infections, are often correlated with the geographical areas where they have been first noted. Most of the recent aggressive flu viruses, such as mexican swine flu, SARS, oriental flu, as well as animal originated ones like chicken virus and swine flu, have been observed in specific geographic areas during initial infection phases, then their evolution have been followed across vast regions or even continents.

It has been shown that environmental pollution (e.g. landfills, especially containing dangerous materials) is responsiblefor the spread of many severe diseases [1, 2, 3, 4]. When studying the evolution of an epidemiological phenomenon over time and space, dealing with a simple statistical model is not sufficient any more. The researcher not only needs to see the status of the current disease, but also to correlate it with other (i.e. similar or related) phenomena occourred in other moments. For example it would be useful to correlate the presence of electromagnetism sources with the occurrence of a particular disease or to see what happens to certain life parameters in a population living in areas near to landfills. Statistical studies about such correlations have been exploited in several research works, as for instance [7] for prostate cancer or [8], where breast cancer data are processed with the help of a GIS application and analyzed using spatio-temporal queries. Nevertheless, none of the current work offers an integrated GIS-based platform for studying and monitoring health related phenomena.

We present a framework for representing and querying health patient information related to geographical information. Patient information is first extracted from digital patient records, loaded in a geographical database and then associated with a position on a geographical map. Points of interest are also stored in the GIS and may be related to patient information for answering spatial queries. Topological queries may be formulated to gather knowledge about possible correlations between diseases diffusion and environmental information.

In this paper we present design and implementation issues of a geographical information system hosting both clinical and biological information. The developed system prototype has been experimented using a large dataset to show its feasibility. Several queries have been performed using both an SQL-like language and a graphical user interface, in order to show the usability of the system and its power in terms of data analysis.

The presented prototype, called *Geomedica*, has been implemented and tested importing patient records of an association of general practitioners, using a common database structure for managing their data. Such a medical doctor community, living in a large area of an Italian region, furnished a test data set of about 200.000 patient records regarding ten years of observed and cured (where possible) diseases. Such data have been provided together with patient residence, thus that records have been anonymized and mapped into the geographical system. An ad-hoc module is in charge of automatically normalize and include data into a core PostgreSQL/PostGIS [13, 14] database instance, just after data have been geocoded.

The paper is organized as follows: Section 1.1 presents related works about geographical information systems applied to health community; Section 2 presents the data modeling, query language, data preprocessing and geocoding used in the Geomedica system; Section 3 presents the system architecture; Section 4 presents the use of the system, with some query examples. Finally Section 5 concludes the paper and foresees future works.

1.1. Related Works

Geographical epidemiology is the science which associates epidemiological phenomena to geographical areas. There have been many works in such a direction, both from Computer Science and from Medical communities. In some cases, the use of GIS with statistical analysis has been proposed including also population distribution, temporal information and diseases information positioning. For instance in [6], several statistical models for pointly represented disease patterns are shown, whereas [7] presents several examples of geographical epidemiology for prostate cancer and [8] for breast cancer. Nevertheless, all the proposed methods study a posteriori disease phenomena by correlating the data with GIS. None of the existing works propose a geographical based framework to formulate queries in charge of extracting and (eventually) foreseeing disease distribution. In [5] a GIS is used to analyze a swine related disease in the geographic distribution, to analyze the distribution, while GIS are used in [12] for drawing up disease maps and for ecological analysis. Other uses of geographical information system have been reported in [9] and [11].

The possibility of using GIS and on line analytical processing (OLAP) techniques for analyzing geographical data has been also reported in [10] where a software prototype for interactively exploring clinical data have been presented.

The use of geographical information systems [11] has been used in several application areas where relating alphanumerical information with geographic map is useful. For instance: (i) *geomarketing* applications associate business information to map locations; (ii) *geogovernment*, where data need to be integrated and mapped to geographical

systems for controlling and monitoring information (e.g., local tax control, technological networks, land use); (iii) *geodefense*, where military resources and personnel (dynamical) location are mapped on geographical databases to plan and control operations, and land disposal; (iv) *geoenvironment*, where data about pollution and climate conditions are analyzed through a geographycal information system; (v) *geomedical*, used to map and study epidemiological phenomena onto geographical system. The here proposed work fits the latter case: there exist many examples of epidemiological studies related to disease distributions related to environmental phenomena or simply related to monitor disease distributions. Nevertheless there is no study on Geo-localization of health data on public web as services able to aggregate information for each community working on health data.

2. Geographical Functions and Medical Data

2.1. Data model

Spatial data models provide the possibility to manage geometric information together with alphanumeric information. In this work we consider a simple spatial model where plain geometrical and topological information such as *point*, *line* and *polyline* are considered. Two dimensional geometric waves, such as *circles*, *curves* and more general *user defined windows* are used in the query evaluation layer, but are not treated in the spatial database. Therefore, the adopted model is based on a logical layer where alphanumerical and simple geometric information are treated as a single layer thematism, whereas at the physical layer a relational database model with spatial extensions is used to store both alphanumerical and geospatial information. Geographical information are added to standard relations by adding a further column specialized in storing geospatial information such as: (*i*) type (e.g. point, line, polyline) and (*ii*) spatial values (e.g. *POINT*(10 33)).
Consider, for instance, the standard relation

```
patient(id, sex, address)
```

The insertion of a further column storing geospatial information can be carried out by executing a stored procedure. For instance, the SEQ statement

```
SELECT AddGeometryColumn('patient', 'loc_geom', -1, 'POINT', 2)
```

where *AddGeometryColumn*() is a stored procedure which is in charge of adding the column `loc_geom` to the table `patient`; its parameters −1, `POINT` and `2` specify, respectively, (*i*) the coordinate system to be used to represent geometry coordinates (-1 means none), (*ii*) the geometric type and (*iii*) the arity required to denote a point (i.e. latitude, longitude or x, y).

Query language

The extended database can be queried by means of extended SQL queries. In this work we refer to the language PostgreSQL extended with PostGIS for spatial queries [14]. PostGIS extends SQL by adding several geometrical and topological functions such as Area, Perimeter, Distance, Containment, Intersection, Union, etc. For instance, the query

```
SELECT count(id)
  FROM patient
 WHERE distance(loc_geom, 'POINT(102030 304050)') < 2500;
```

computes the number of patients who live within a fixed distance from a given point.
It is even possible to define more general queries using geometrical functions. For instance, the query computing *n*umber of patients in a user designed polygon can be expressed as follows:

```
SELECT count(id)
  FROM patient
 WHERE ST_Contains(loc_geom, 'POLYGON(0 0,0 1000,1000 1000,1000 0,0 0)')
```

where `POLYGON` defines a polygon from a set of points and `ST_Contains` is a boolean function in charge of selecting the points associated with $\texttt{loc}_{\texttt{g}}\texttt{eom}$ which are contained in the polygon.

2.2. Data preprocessing

Before relating data with spatial information, alphanumeric information are normalized to reduce inconsistency. This is carried out by a number of scripts which during the normalization phase support the user in discovering and correct inconsistent information. For instance, the below script, replaces the shortcuts P.ZA, P.ZZA, P ZA with PIAZZA (the italian name for square) and the shortcuts C.SO, C.RSO, CSO with CORSO (the Italian name for avenue).

```
cat $1 |awk -F";" '{ print $3,"," $5, $4 }' | sed 's/"//g' |
sed 's/P.ZA/PIAZZA/g' | sed 's/P.ZZA/PIAZZA/g' | sed 's/P\/ZA/PIAZZA/g' |
sed 's/C.SO/CORSO/g' | sed 's/C.RSO/CORSO/g' | sed 's/CSO/CORSO/g' ...
```

It is worth noting that the full names used in the replacement are the same used by the geo-referred maps.

Identification of inconsistency can be carried out in a semi-automated fashion by using integrity constraints (i.e. it is possible to write queries showing pairs of inconsistent tuples with respect to some functional dependency). Let's assume, for instance, the functional dependency id → address to be true for the relation patient, meaning that two patients with the same id must be associated with the same address. In order to discover pairs of tuples which do not satisfy the functional dependency, the following query can be executed:

```
SELECT P1.id, P2.id
  FROM patient P1, patient P2
 WHERE P1.id = P2.id AND P1.address <> P2.address
```

2.3. Georeferencing

Data geocoding, that is the association of alphanumerical and geospatial information, has been realized by using the Google Map APIs.
The script reported below associates each patient to a position on the map by using its alphanumeric address.

```
private String googleRetrieve(String addr) throws GeocodingException {
  // transform the address in UTF-8 format
  String address = URLEncoder.encode(addr, "UTF-8");
  // create the URL object
  URL page = createFinalURL(GOOGLE_ADDRESS, finalAddress);
  // open a connection and retrieve the coordinates
  [...]
  return retrievedCoordinates;
}
```

The result is then saved in the loc_geom attribute of the patient relation.
The choice of Google Map is motivated by the fact that it is free for research purposes, contains reliable and up to date information and provides sophisticated functionalities to interactively navigate through query results [15].

3. Architecture

Geomedica is a cartographic system designed for mapping health information into a geographical database and working as a web application. The system has been implemented on the Google map service and uses the Google map APIs for representing, manipulating and visualizing geographical data. Using a loading module, Geomedica is able to import alphanumerical data that are geographically mapped into the Google Map cartography. Finally, the system prototype provides a graphical user interface to formulate alphanumerical and spatio-temporal queries.

We have experimented the system prototype with a large dataset containing clinical data produced from an association of medical family doctors. Moreover, we have enriched this dataset providing point of interests such as electrical central power, electromagnetic field and landfills, which are potentially associable to healthy information.

The architecture of the system, designed for the monitoring epidemic phenomena, is reported in Figure 1. It consists of the following modules:

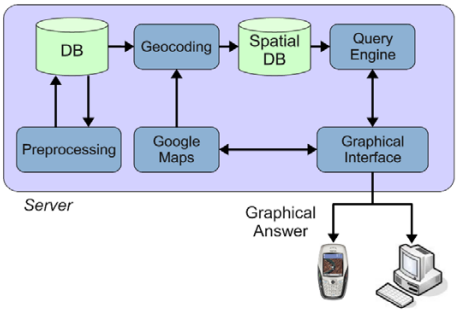


Figure 1: System Architecture

- *Preprocessing*, that is in charge of normalizing data. This module gives also support to integrate different data sources, as well as to clean and correct incorrect and inconsistent data. Data integration techniques are necessary as often source data are of different types (relational tables, flat files, semistructured data) or are stored into relations with different schema. Since in the context of clinical data privacy preserving is mandatory, this module also performs anonymization of data by removing/rewriting sensible personal data (i.e. name, surname, etc.);

- *Geocoding*, which is in charge of translating alphanumerical patient addresses into geographical locations (i.e. latitudes and longitudes) in order to geocode patient information and store it in the *Spatial DB*;

- *Google Maps*, a wrapper for Google Maps API [15]; this module provides facilities to invoke the Google Maps geocoding engine, retrieve results and manage geocoding errors (i.e. malformed or non-existing addresses); furthermore it is also called by the *Geographical Interface* module in order to visualize spatial data on the map;

- *Query Engine*, a module not only devoted to the computation of query results (calling the PostGIS query executor), but also designed for interpreting data filters specified by the user via the graphical user interface and translating data from raw database resultsets to Java object collections, making them ready for map rendering;

- *Graphical Interface*, a module designed for the end user to easily submit queries; this module has been also designed to render georeferenced objects and alphanumerical data contained in database query results (obtained from the *Query Engine* module) on the client's device; it adapts information to make it readable and browseable according to specific client-side limitations if any (e.g. smart phones, clients with limited resources); mapping projections and functions are obtained invoking the *Google Maps* module.

It is worth noting that in the architecture we can identify a number of layers which communicate so that the information can be extracted, processed and correctly rendered in the desired format. Moreover, in the logical architecture there are also two databases:

- an alphanumeric database, containing the epidemiologic and clinical information about patients, extracted and unified from heterogeneous data sources (i.e. public health structures, hospitals, general pratictioners);

- a spatial database, storing and representing geographical information to be merged with clinical (alphanumeric) one; it contains data on digital cartography, patient locations, POIs (points of interest), area boundaries, etc.; these data can be gathered from specific companies (i.e. Navteq, TeleAtlas) or used through some cartographic API (i.e. Google Map, in our case).

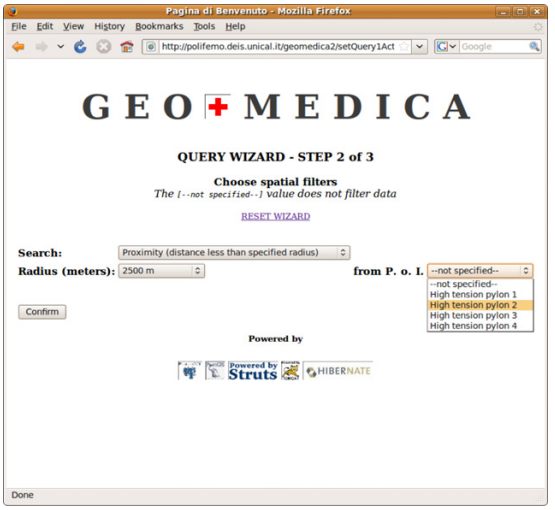


Figure 2: Geomedica web interface: Spatial constraints definition. The user, at each step of the web wizard, can specify a number of contraints to filter query results. During the above step, a simple spatial contraint is shown (i.e. radius-based query).

4. Using Geomedica

In this section we show how spatio-temporal queries can be expressed using the web interface provided by Geomedica.

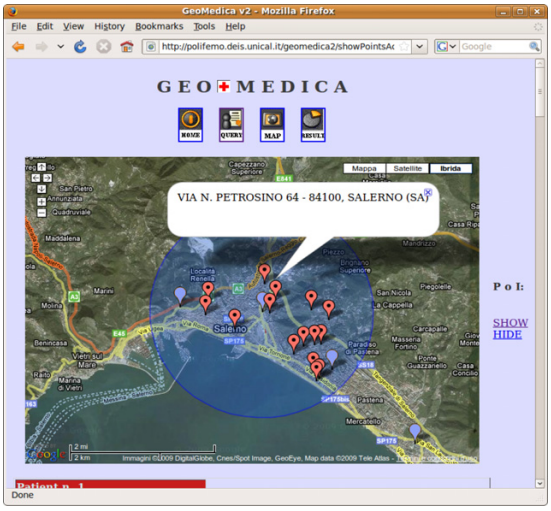
Selecting the survey analysis phase, the user is guided through a wizard for query composition.
In the first step of the Geomedica application the user can specify filters over the alphanumeric dataset attributes. In such an example, we look for diseases in a user defined interval of the international *icd9_cm* code for disease specifications, and other alphanumerical information such as sex and age. In a further step, by specifying some geographical point of interest, the query may be enriched by using topological constraints. For instance, we may look for patients within a given area (containing a point of interest such as an high voltage pylon) or within a given distance from a POI (see Figure 2).

The query will be evaluated on the server side over the geographical database and the results will be shown in a cartographic map, powered by Google map (see Figure 3(a)), where all the Google functionalities are shown (zoom in, zoom out, panning, and layer visions). Moreover, data results are represented with special pinpoints, that are associated with records in the database, and also with points of interest. The query results also return tabular format of filtered information. Figure 3(b) shows the tabular information that is reported in the lower part of the web page result. Note that (*i*) only patients that have in their clinical history the filtered values of the *icd9_cm* disease (i.e. value in the range of $[100 - 120]$) are reported and (*ii*) the complete clinical history of patients is showed in the result.
It is important to observe that the system allows us to also express queries using the SQL language and that the user interface has been designed for end users.

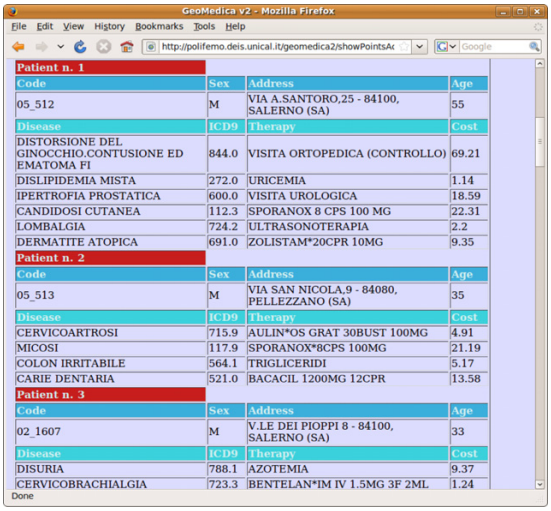
Geomedica system has been tested with several predefined queries, suggested by domain experts, such as

1. Select all patients affected by a given disease and show results with variable colors associated with density regions (e.g., such as number of cases for area);
2. For each point of interest and for each disease, compute the number of patients, within a given distance, affected by the disease.

Furthermore, the system allows us to interactively add new points and areas of interest and to load other datasets for more complex tasks (e.g. we have also used datasets of pharmacies, for monitoring drug expenses).



(a) Mapped results



(b) Tabular results

Figure 3: Geomedica web interface: interactive geographical map representing data on the left and tabular view of results

Results elaborated by the Geomedica system are receiving good feedbacks from domain experts (physicians and medical doctors), who say they are very useful. Thus we are now planning to integrate data mining techniques in the system to help them in discovering hidden relations among diseases and geographical areas, to define and test previsional models (e.g. linking ages and areas with the most likely diseases).

5. Conclusion and Future Works

The Geomedica platform is a system prototype implemented to efficiently analyze and visualize datasets regarding health information and can be used for both epidemiological studies and for land and spatio-temporal analysis in Health. It has been showed that the correlation of alphanumeric and spatio-temporal information and the efficient indexing improve the decision support task in several business areas (e.g., marketing analysis, education and evaluation, industry, military planning).

We are currently performing new tests with two new large datasets. The first one is about alcohol consumption and alhoholism; the information contained in the dataset is obtained from associations, hospitals and military forces and is about the distribution and cases of alcohol uses, as well as feedbacks of a wide spread anti-alcohol campaign. This kind of application has already received great interests as alcohol is one of the most important cause of death in car accidents. The second dataset concerns the screening of population with respect to TSA values, that is information about hypothyroidism pathology. Similarly, information needs to be associated with particular areas contaminated by pollution, as well as areas where population has changed the use of salt in their diet or areas nearby cost lines.

Acknowledgements

We would like to thank the association of general doctors of the Campania Region (Italy) for providing the dataset and for their helpful comments on the system.

References

[1] C. A. Pope III, R. T. Burnett, G. D. Thurston, M. J. Thun, E. E. Calle, D. Krewski, J. J. Godleski, Cardiovascular Mortality and Long-Term Exposure to Particulate Air Pollution, *Circulation*, 109, pp. 71-77, 2004

[2] U. Ackermann-Liebrich, P. Leuenberger, J. Schwartz, C. Schindler, C. Monn, G. Bolognini, J.P. Bongard, O. Brandli, G. Domenighetti, S. Elsasser, L. Grize, W. Karrer, R. Keller, H. Keller-Wossidlo, N. Kunzli, B.W. Martin, T.C. Medici, A.P. Perruchoud, M.H. Schoni, J.M. Tschopp, B. Villiger, B. Wuthrich, J.P. Zellweger and E. Zemp, Lung function and long term exposure to air pollutants in Switzerland. Study on Air Pollution and Lung Diseases in Adults, *American Journal of Respiratory and Critical Care Medicine*, 155(1), pp. 122-129, 1997

[3] M. Harada, Minamata Disease: Methylmercury Poisoning in Japan Caused by Environmental Pollution, *Critical Reviews in Toxicology*, 25(1), pp. 1-24, 1995

[4] K. R. Smith, S. Mehta, The burden of disease from indoor air pollution in developing countries: comparison of estimates, *International Journal of Hygiene and Environmental Health*, 206(4-5), pp. 279-289, 2003

[5] Z. Matzkin, S. Alexander, M. Brumsted, J.T. Shissler, T.D. Parsons, Geographical Information System (GIS) Mapping of swine disease dynamics, *International Symposium on Emerging and Re-emerging Pig Diseases*, Rome, 2003

[6] A. C. Gatrell, T. C. Bailey, P. J. Diggle, B. S. Rowlingson, Spatial Point Pattern Analysis and Its Application in Geographical Epidemiology, *Transactions of the Institute of British Geographers*, New Series, 21(1), pp. 256-27, 1996

[7] L. Jarup, N. Best, M. B. Toledano, J. Wakefield, P. Elliott, Geographical epidemiology of prostate cancer in Great Britain, *International Journal of Cancer*, 97, 2001

[8] V. M. Vieira, T. F. Webster, J. M. Weinberg, A. Aschengrau, Spatial-temporal analysis of breast cancer in upper Cape Cod, Massachusetts, *International Journal of Health Geographics*, pp. 7-46, 2008

[9] M. Kwan, Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set, *Transportation Research Part C: Emerging Technologies*, 8(1-6), pp. 185-203, 2000

[10] Y. Bedard, P. Gosselin, S. Rivest, M. J. Prolux, M. Nadeau, G. Lebel, M. F. Gagnon, Integrating GIS components with knowledge discovery technology for environmental health decision support, *Int. Journal of Medical Informatics*, 70, 2003

[11] J. R. Meliker, M. J. Slotnick, G. A. Ruskin, A. Kaufmann, G. M. Jacquez, J. O. Nriagu, Improving exposure assessment in environmental epidemiology: Application of spatio-temporal visualization tools, *Journal of Geographical Systems*, Springer Berlin / Heidelberg, 7(1), 2005

[12] T. Kisteman, F. Dangendorf, J. Schweikart, New perspective on the use of Geographical Information Systems (GIS) in environmental health science, *Int. J. Hyg. Environ. Health*, 205, pp. 169-181, 2002

[13] M. Stonebraker, L. A. Rowe, The design of POSTGRES, *ACM SIGMOD Record*, 15(2), pp. 340-355, 1986

[14] S. Holl, H. Plum, PostGIS, *GeoInformatics*, 3, pp. 34-36, 2009

[15] Google Maps Website, *http://maps.google.com*