Studying Neonatal TSH distribution by using GIS

Giuseppe Tradigo  
DEIS - Electronics, Computer & System Sciences Dept.  
Univ. of Calabria - Rende, Italy  
gtradigo@deis.unical.it

Pierangelo Veltri  
Surg. & Medical Science Dept.  
Univ. of Catanzaro  
veltri@unicz.it

Onorina Marasco  
Neonatal Screening Center  
Materdomini Polyclinic  
neonatalscreening@unicz.it

Giovanna Scozzafava  
Neonatal Screening Center  
Materdomini Polyclinic  
neonatalscreening@unicz.it

Giuseppe Parlato  
Health Science Dept.  
Univ. of Catanzaro & Neonatal Screening Center  
Materdomini Polyclinic  
parlato@unicz.it

Sergio Greco  
DEIS - Electronics, Computer & System Sciences Dept.  
Univ. of Calabria - Rende, Italy  
greco@deis.unical.it

ABSTRACT
Geographical Information Systems, i.e. GIS, are used to help communities in managing data related to their geographical location. Associating textual data with spatial extension and time can be crucial to understand and improve human health. Exploiting available data and extracting new knowledge can lead to disease distribution and migration models (e.g., epidemiology).

In this paper we report the experience of using GIS technologies to analyze clinical data containing TSH values about newborn in a spatially delimited region. TSH neonatal screening has been performed on blood of newborn with the aim to discover diseases at an early stage and to study the detect any possible arise of hypothyroidism. We present a flexible geographical system called Geomedica which is being used to analyze such data according to a two steps approach: (i) study of the last 10 years of data distribution in an Italian region with over 18 thousands newborn per year and (ii) identify possible clusters by querying and projecting results geometry on a thematic map.

Queries performed on the available dataset were able to correctly correlate health data about patients with geographical features (e.g., points of interest, boundaries, coastline vectors) and to visualize diseases distributions on a geographical map. However, the proposed queries may be considered as an important starting point for similar environment-dependent pathologies.

Categories and Subject Descriptors
H.2.8 [Database Management]: Database Applications—Spatial databases and GIS; J.3 [Life and Medical Sciences]: Health

General Terms
Experimentation, Measurement

Keywords
GIS, human health, epidemiology

1. INTRODUCTION
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. It has been proved that associating data with (also time varying) spatial position, can be used to study factors or improve knowledge related to information, habits or to discover information. Applications can be used to associate geographic information with data, obtaining trajectory and patterns. Such a strategy can be definitely applied to study human health. A GIS with alpha-numerical health data allows, for instance, to study the environmental influence on human life, as well as diffusion of pathologies in space and time (e.g. flu virus distribution and migration). Thus, storing and querying health related data together with geographical information, allow to study and better understand health related phenomena. However, statistical analysis and mining techniques can be used to extract disease models or to predict disease diffusion areas. Epidemiological phenomena, as well as new virus infections, can be correlated with the geographical areas where they have been first noted. For instance, annual flu viruses are studied with respect to their geographical origins. Moreover environmental factors can be associated to diseases and, more generally, with human life quality. For instance, many researchers are working on finding correlations between environment and cancer disease and deaths: e.g. causes such as high tension electric power lines, asbestos in buildings, or industrial areas with high pollution productions, have been studied as high risk factors. Geographical information is useful while studying any phenomena or behaviour where geographical location improves data analysis and decision taking. This is very useful when studying diseases and their evolution, concentration, or any epidemiology phenomena.

Neonatal TSH screening, is an analysis performed on newborn children blood, and aims to identify the Thyroid Stim-
ulating Hormone (TSH), helping the early detection of thyroid diseases. The Neonatal screening Department, which provided the dataset for the here presented project, also treats neonatal blood samples in order to discover other diseases. However, here we just consider geographical and environmental influences on the TSH measured values. Studying the TSH consists in comparing the TSH measured for a newborn, with a reference value. The reference value, considered also as a disease threshold, is obtained by considering the TSH values measured for entire population in a certain period of time (e.g. last year) and calculating the 99th percentile. All the newborn presenting a TSH value over this threshold, are selected for further analysis and diagnosis. The screening is performed on the whole children population, thus that the TSH value trend can be studied with respect to residential information. However, what experts are now coming up to understand, is that normal TSH value ranges are influenced by environmental and behavioural factors. In this paper, we present a project on analyzing the TSH values from neonatal screening performed in a South Italian region, Calabria, in the last 10 years.

2. RELATED WORKS

There have been many works associating geographical data to health data, 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. In [3], several statistical models to represent disease patterns are shown, whereas [4] presents several examples of geographical epidemiology for prostate cancer and [5] 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 (possibly) foreseeing disease distribution. The idea here presented is to associate the TSH distribution with the geographical information trying to improve the reference threshold and to have reliable information about clinical data. In [2] a GIS is used to analyze a swine related disease according to a geographic distribution, while in [8] GIS are used to draw disease maps and for ecological analysis. Other uses of geographical information system have been reported in [6] and [7].

Geographical information systems have 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 geographical 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. Moreover, in the here presented project, the entire population of newborn children is taken into account thus representing a screening example and not only an analysis of disease cases. Even if the idea of considering healthy cases distributions in the Calabria region is not new (see for instance the SigMCC project for cancer data [12]) there is no example of mapping such information into geographic areas.

3. USING GEOMEDICA FOR THE TSH APPLICATION

The Geomedica dataset has been anonymously obtained from the Neonatal screening center of Mater Domini Hospital of Catanzaro, which performs around 20 thousands tests per year. This dataset contains data collected since 2005. The 99th percentile of the whole dataset varies from 1.5 to the current 0.7 value in the last 7 years. This means that there is a measurable variation in the newborn TSH measured values. However, even if the disease threshold has been constantly decreasing over time, the number of ill subjects increases. Studies for such a phenomenon seem to relate the variation to the newborn parent’s behaviour and to environmental factors. Indeed, different distributions have been observed for different cities for the Calabria region. The geographical positioning of the newborn children is related to the position where parents lived during the 9 months of gestation. The aim of this project is to correlate TSH values with geographical residential information of newborn parents to study and possibly identify environmental influences as a cause of TSH anomalies. Moreover, as a possible target, the project aims to discover significant differences in identified TSH value clusters which could even be extended toward the understanding of other diseases. This can be done by: (i) inserting the geographical and environment information into the evaluation process of the TSH normal threshold and (ii) identifying possible clusters of attention for geographic area where environmental anomalies may influence the thyroid healthy status. For example, preliminary tests show that part of the newborn children (almost 30 %), mostly born close to the sea, have a TSH value that is above the disease threshold. These values are concentrated in particular areas, meaning that probably there are many false hits due to the fact that the considered 99th percentile (disease threshold) should be probably tuned according to attributes of the geographical area (e.g. altitude, distance from the sea coast).

This project uses an open source GIS called Geomedica which has been developed in the Bioinformatics Laboratory of University Magna Graecia. The GIS framework allows to map health patient data (in this case newborn children residential and blood analysis data), into a geographical database system. Geomedica offers functionalities to import patient data, to geocode residential information and to store it in a PostgreSQL database [9, 10]. The user interface has been designed to be web oriented so that it can be used with a web browser to pervasively formulate queries and visualize results. The system has been tested with clinical data for different health applications (see [1]). For a given database instance, Geomedica is able to extract information about clusters and define relevant areas for values of interest (in this application TSH values). Topological
queries may also be formulated to gather knowledge about possible correlations between diseases diffusion and environmental information. Finally, due to the influence of altitude to TSH, we plan to include altitude information on the geographical area to associate ionization level of the newborn parent residential with respect to TSH measured values.

The system has been implemented on the Google maps service and uses the Google maps 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.

The architecture of the system, is reported in Figure 1. It consists of the following modules:

- **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. 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; in the current application it is in charge of associating geographical data to neonatal TSH values.

- **Google Maps**, a wrapper for Google Maps API [11]; this module provides facilities to invoke the Google Maps geocoding engine, retrieve results and manage geocoding errors (i.e. malformed or non-existing addresses); in our case it also associates altitude information;

- **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 designed to render georeferenced objects and alphanumerical data contained in database query results (obtained from the Query Engine module) on the client’s device;

Neonatal alphanumerical schema is reported in the following: NeonatalData(NationalCenter, ResidentialData, Address, Sex, DateOfBirth, weight, MotherThyroid, TSHValue).

The schema has been enriched with geographical attributes (i.e. latitude, longitude, altitude).

### 4. EXPERIMENTS AND PRELIMINARY RESULTS

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 health information. Thanks to a flexible user interface, and to a modular query engine, Geomedica offers the possibility to design queries in a simplified way, with a module devote to the queries by example.

Preliminary results on analyzing TSH data distribution allowed to identify a much more correct TSH threshold related to Calabria region. The TSH threshold evaluated on the whole population was not exactly covering the population distribution. Indeed, in the northern part of the same region, 35% of the newborn population has a 99th percentile TSH disease threshold lower than the southern part. This is due to the fact that the southern part the population lives closest to the sea coast, where the iodium concentration is usually higher due to the sea fumes. In such a case, for example, a higher 99th percentile should be considered for the southern part with respect to the northern. Moreover the geographical information allowed us to analyze data by using geometrical features (e.g. query boundaries). Indeed, residential data are associated to the 5 main cities (and thus counties related) of the region, where counties have a non regular land distribution. Thus newborn belonging to two different cities, may be much more close to common interest points (e.g. an electromagnetic source) and captured if selected by drawing geometric region (e.g. circle, rectangle) on the graphical user interface.

In Figure 2 the distribution of annual newborn in two cities of the Calabria region (southern part of Italy) is reported. In figure and a selection query based on TSH values above a threshold inside a circled area is shown. Both scenarios allow the final user to interact the query result geometry and explore map regions at different detail level, adding or hiding other information layers.

### 5. CONCLUSION

We expect to obtain interesting results and guiding choices about TSH value threshold and ranges. Another interesting outcome will be clusters and concentration patterns for suspicious TSH values observed in the regions of interest.
6. ACKNOWLEDGMENTS

This paper is partially supported by the CARDIOTECH project, funded by the Italian Research Ministry.

7. REFERENCES

[1] G. Tradigo, S. Greco, P. Veltri Geomedica: managing and querying clinical data distributions on geographical database systems IEEE ICCS 2010,


