Epidemiological Data Analysis in TerraFly Geo-Spatial Cloud

Huibo Wang, Yun Lu, Yudong Guang, Erik Edrosa, Mingjin Zhang, Raul Camarca, Yelena Yesha, Tajana Lucic, Naphtali Rishe

NSF Industry-University Cooperative Research Centers and
School of Computing and Information Sciences
Florida International University
Miami, Florida, 33199
{ huibwang, yun, yuguang, eedro001, zhangm, rcamarca, rishe}@cs.fiu.edu,
yeyesha@csee.umbc.edu, drlucic@gmail.com

***Abstract*. GIS systems and online services are growing at a very fast pace; however, there are few online services for the analysis of geospatial epidemiology and their functionality is limited. We present a geospatial epidemiology analysis system on the TerraFly Geo-spatial Cloud platform. The system provides comprehensive spatial analysis methods and visualization. In this system, the user is not required to program in order to employ the functionality. All the datasets are stored in the Geo-spatial Cloud.**

This system is accessible at http://terrafly.fiu.edu/GeoCloud/. The system API algorithms adapted to geospatial epidemiology. The application utilizes the GeoCloud distributed storage system for the Big Data to be analyzed; it utilizes an interactive mapping API to display results.

Keywords: geospatial epidemiology, GIS; visualization, Big Data

I. INTRODUCTION

Nowadays, GIS systems grow at a very fast pace, and much of functionality has moved to online services. Geospatial epidemiology is a discipline which benefits from geospatial analysis, but there are few online services suitable for geospatial epidemiology, and their functionality is quite limited. If spatial epidemiology is enabled in online GIS systems to provide an online, easy-to-use analysis and visualization service, then this would be a useful tool to epidemiologists, as well as to users interested in searching disease maps.

Epidemiology employs a large number of analysis methods that are not easily used without programming. Furthermore, epidemiological analysis requires data related to other disciplines, such as economics, ecology, and demographics, which are not easy to retrieve and employ in the analysis. The system should display results on an interactive map and have the capability to store and compute Big Data.

We present here a spatial epidemiological analysis system based on the TerraFly GeoCloud. This system can provide online analysis with comprehensive main algorithms designed especially for geospatial epidemiology, which are easy to be used without any programming knowledge. Also, users can employ in the analysis various related datasets from the TerraFly GeoCloud system, such as socioeconomic and demographic data. Further, this system allows users to visualize results on a map. This system can provide distributed storage, allowing users to upload, manipulate, and download data.

This geospatial epidemiology analysis system is based on the TerraFly GeoCloud. It makes use of the TerraFly GeoCloud storage and visualization API, as well as of the Django web framework. The algorithms are implemented using the Python and R languages. Compared with other services and software, this system can provide easy-to use, comprehensive, and friendly spatial epidemiology analysis methods and clear visualization for the benefit of both the spatial epidemiology researcher and of the common user concerned about public health. The system is available at http://terrafly.fiu.edu/GeoCloud/ .

The rest of this paper is organized as follows. In Section 2 we present the background of this system. In Section 3 we describe the architecture of the epidemiology analysis system within TerraFly GeoCloud. In Section 4 we present the analysis algorithms and visualization solutions. In Section 5 we present a case study based of lung cancer mortality in Florida, in order to show how to use this system to perform epidemiological analysis. In Section 6 we discuss related software and online services.

II. BACKGROUND

A. *Geospatial Epidemiology*

Geospatial epidemiology is a field of study focused on describing and analyzing the geographic variations of disease spread. Specifically, geospatial epidemiology considers demographic, environmental, behavioral, socioeconomic, genetic, and infectious risk factors [1]. Geospatial epidemiology has three focus areas: 1. disease mapping, 2. geographic correlation studies, and 3. clustering, disease clusters, and surveillance.

Recent advances in data availability, GIS systems, and analytic methods have created new opportunities for investigators to improve the display of geospatial epidemiology analysis results.

Disease maps, which provide a direct visual summary of complex geographic and disease information, have a long history. In recent years, disease mapping as a fundamental part of disease analysis has attracted a considerable amount of effort. Disease maps typically include mortality maps, morbidity maps, standardized mortality/morbidity ratio (SMR) maps, age-adjusted ratio maps, etc. The SMR ratio is often shown for geographic areas such as countries, counties, zip codes, or other granularities. This information guides in understanding which areas have a high risk of disease contagion [6].

The aim of geographic correlation studies is understand the effect of various factors on the spread of diseases. These factors may relate to geography, socioeconomics, demographics, or lifestyle. For example, is lung cancer is related to smoke, or does income affect the risk of a given disease [3][5].

Disease cluster detection attempts to discover those groups that are at higher or lower risk of disease transmission. By analyzing these clusters, researchers can also discover relative factors of the disease. For example, consider people living in a heavy pollution area; they may be subject to a higher disease risk, and they form a cluster that can be detected and analyzed. Existing disease cluster detection methods include Local Moran's I[15], Scan statistic[7] and Getis-Ord Gi* statistic[12].

Disease surveillance aims to monitor the spread of disease and to report disease outbreaks.

B. Spatial data visualization

Visualization of numerical and symbolic data helps create a clearer understanding of the data. Visualizations may also provide new ways to interpret data by enabling researchers to recognize patterns. Data can be visualized in many ways, including charts, graphs, images, and tables. Spatial data, or geographic data, contains coordinates, allowing their position in space to be mapped. Coordinates usually represent a position in a two- or three-dimensional plane. Other data fields can then be used in various visualization techniques that display and change the appearance of the mapped spatial data, allowing a quicker understanding and comparison of the different data values. In geospatial epidemiology, visualization of data and of analysis results is important in helping researchers understand the spatial relationships between different areas.

C. TerraFly GeoCloud

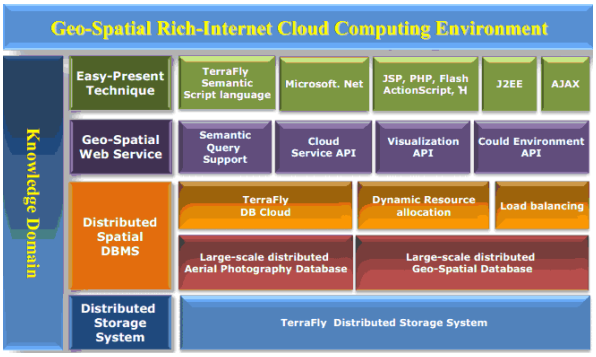

Fig. 1. The Architecture of TerraFly GeoCloud

TerraFly GeoCloud[2] is a system combining several diverse technologies and components in order to analyze and visualize geospatial data. Figure 1 depicts the architecture of TerraFly GeoCloud. In this system, a user can upload a spatial dataset and display it using the TerraFly Map API[4]. Datasets can be subsequently analyzed using various functions, such as Kriging, a geo-statistical estimator for unobserved locations, and spatial clustering, which involves the grouping of closely related spatial objects. The results of these spatial analysis functions can then be displayed in charts or other map visualizations. The system is preloaded with several sample data sources, including include demographic census, real estate, disaster data, hydrology, retail, and crime data. The system also supports MapQL, a query language that features an SQL-like syntax to create custom map visualizations.

Various analysis functions related to spatial epidemiology have been integrated into TerraFly GeoCloud. Analysis functions can be used by selecting the appropriate dataset and function in the menu, and selecting the variables to be analyzed. TerraFly GeoCloud then processes the data and returns a result that can be visualized on the TerraFly Map, or on a chart. Results displayed on the map include a legend, which identifies certain range values by color. Certain visualizations are interactive, allowing additional information to be displayed.

III. SPATIAL EPIDEMIOLOGY ANALYSIS AND VISUALIZATION SYSTEM BASED ON TERRAFLY GEOCLOUD

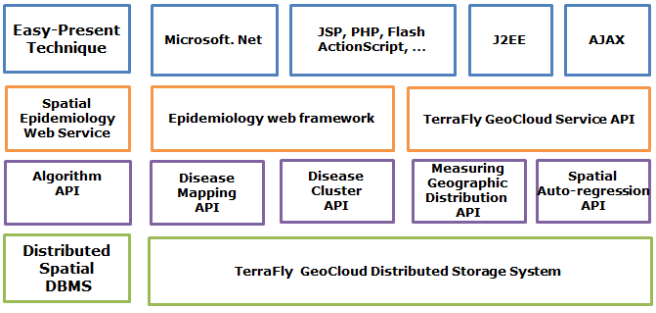

Fig. 2. The Architecture of the Spatial Epidemiology System

The architecture of the spatial epidemiology system is shown in Fig. 2. This system, based on TerraFly GeoCloud,

provides online spatial epidemiology analysis and visualization services. It makes use of the TerraFly GeoCloud Distributed System for data storage and of the TerraFly GeoCloud Service API in order to provide web services. Through them, users can upload and use datasets, draw maps, and customize spatial data visualizations. GeoCloud contains demographic census and socioeconomic datasets that can be incorporated to the epidemiology analysis.

Our Spatial Epidemiology System provides four kinds of API algorithms for data analysis and results visualization, based on the TerraFly GeoCloud System: disease mapping (mortality/morbidity map, SMR map), disease cluster determination (spatial cluster, HotSpot analysis tool, cluster and outlier analysis), geographic distribution measurement (mean central, median central, standard distance, distributional trends), and regression (linear regression, spatial auto-regression).

The epidemiology web framework can be used to perform data interaction, and the TerraFly GeoCloud API provides a spatial data visualization service to display the results. Web techniques are implemented to provide interaction to the user.

IV. Analysis and Visualization of Spatial Epidemiology

A. *Disease mapping*

Disease mapping can display disease spread information on a map and provide a direct summary to users. Disease maps include mortality/morbidity maps, SMR maps, etc. Users can see on maps which areas have higher incidence or mortality. Our geospatial epidemiology system provides a mortality/morbidity map and an SMR map.

SMR, the standardized mortality ratio, is a ratio or percentage measuring the increase or decrease of mortality in the study subareas with respect to the wide areas. SMR can be calculated as follows:

$$SMR_i = \frac{O_i}{E_i} \quad (1)$$

where O_i is the number of observed disease mortality cases in subarea i, and E_i is the expected number of disease mortality cases in subarea i E_i is defined as:

$$E_i = \frac{P_i * O_A}{P_A} \quad (2)$$

where P_i is the population in subarea i, O_A is the total number of observed disease mortality cases in the whole area, and P_A is the total population in the whole area.

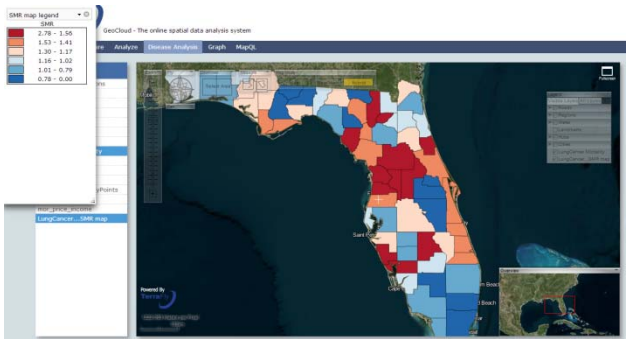

Fig. 3. Standardized Mortality Ratio map

Figure 3 shows a lung cancer SMR map in Florida. Each county has its own SMR. If SMR is close to 1, then the mortality of lung cancer in this county is close to the average mortality in Florida. From this map, users can know which county has lung cancer mortality higher than the average.

B. *Clusters*

Disease Clusters are a common occurrence. Clusters are abnormal aggregations of disease cases in the same period and in the same place. Discovery and analysis of these clusters can help users understand the disease spread mechanism and control the spread of infectious diseases. Disease cluster detection methods include Openshaw's GAM, Besag & Newell, Kulldorff & Nagarwalla, and others.

Openshaw's GAM [11] needs the user to input a radius as argument; the method then scans the area within the circle of the designated radius. The shortfalls of this method include: 1) circles of the same size can refer to different-sized population; 2) it does not account for multiple testing. Besag & Newell [8] needs the user to input the size of the cluster - this is a disadvantage. Kulldorff & Nagarwalla (KN) [7] provides a more accurate method to perform disease cluster detection. The KN method is implemented for circular zones of variable size, and allows for likelihood ratio test. In our geospatial epidemiology system, the Kulldorff & Nagarwalla (KN) technique was chosen to perform disease cluster detection.

The steps of the KN method include: 1. Move a circle in space to obtain an infinite number of overlapping circles; 2. Compute LLR (Log Likelihood Ratio) of each circle and sort the LLRs; 3. Obtain a large LLR, then use the Monte Carlo method to calculate their P-value. The Log Likelihood Ratio can be calculated as follows:

$$LLR = max_j \left(\frac{Y_j}{E_j}\right)^{Y_j} \left(\frac{Y_+ - Y_j}{Y_+ - E_j}\right)^{Y_+ - Y_j} I(Y_j > E_j) \quad (3)$$

where Y_j denotes the observed number of deaths in the circle subarea, Y_+ denotes the number of deaths in the entire area, and E_j denotes the expected number of deaths in circle subarea.

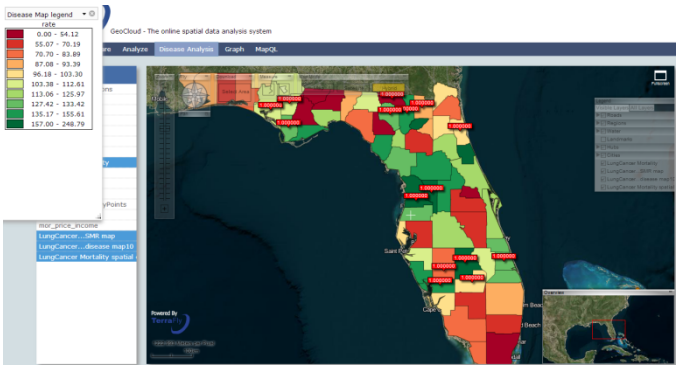

Fig. 4. Disease cluster map

Figure 4 shows a lung cancer cluster map in Florida. The red points indicate disease clusters where unusual disease cases were detected. The number in the red point is the p-value of each area.

Besides the disease cluster, the system also provides HotSpot analysis function, and Cluster and Outlier Analysis.

C. *Regression*

Secondary dataset correlation may be used in epidemiologic analysis, for example: as socioeconomic data. Regression tools provide employ regression models to estimate the relationships between disease data and secondary datasets, such as socioeconomic data. [10]

Regression models include linear regression and spatial auto regression. In linear regression, Moran's I Test is added to test whether the dataset contains spatial correlation. In spatial auto-regression, a lag model and an error model are provided. The spatial auto-regression lag model can be calculated as follows:

$$Y = \rho Wy + x\beta + \varepsilon \tag{4}$$

where Y is a dependent variable, W is a matrix of spatial weights, x is an independent variable, β denotes the unknown parameters, and ε is an error term.

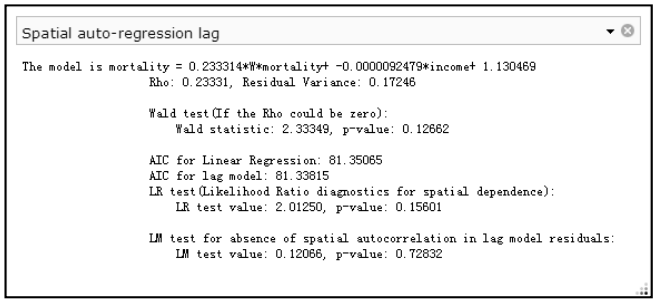

Fig. 5. Spatial auto-regression lag model

Figure 5 shows the result of a spatial auto-regression lag model. In this model, multiple test methods are provided for verifiability: Wald test to determine whether various parameters can be zero or not; AIC for linear regression and lag model, to indicate which model is better; LR test, the Likelihood Ratio diagnostics, for spatial dependence; LM test, for absence of spatial autocorrelation in lag model residuals [13][14].

D. *Measuring Geographic Distribution*

Geographic distribution measurements include mean/median central, standard distance, and distributional trends functions. In our system, a weighted mean central is provided as follow:

$$X = \frac{\sum_i w_i x_i}{\sum_i w_i} \quad Y = \frac{\sum_i w_i y_i}{\sum_i w_i} \tag{5}$$

where x_i and y_i denote the coordinate of each point (but when the data set is polygonal, x_i and y_i indicate the center of each polygon) and w_i is the weight - which corresponds in our system to mortality or incidence.

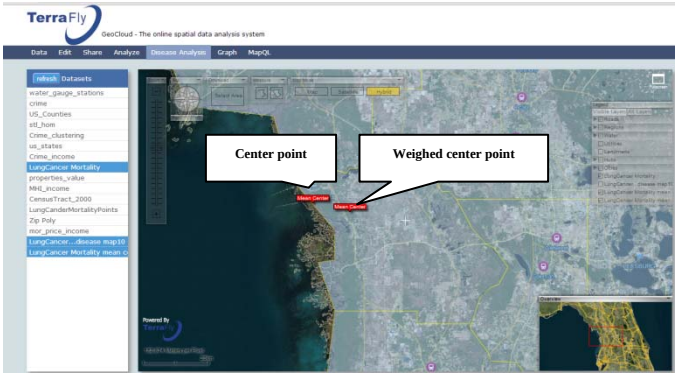

Fig. 6. Center point and weighted center point

Figure 6 shows these two points: one is the non-weighted center point, and the other is the lung cancer mortality weighed center point. They do not coincide.

V. A CASE STUDY

In this section we provide an example of how our geospatial epidemiology system can be employed in epidemiologic research. Assume a researcher studies lung cancer in Florida. She can upload and choose the mor_price_income dataset to TerraFly GeoCloud - shown in Figure 7.

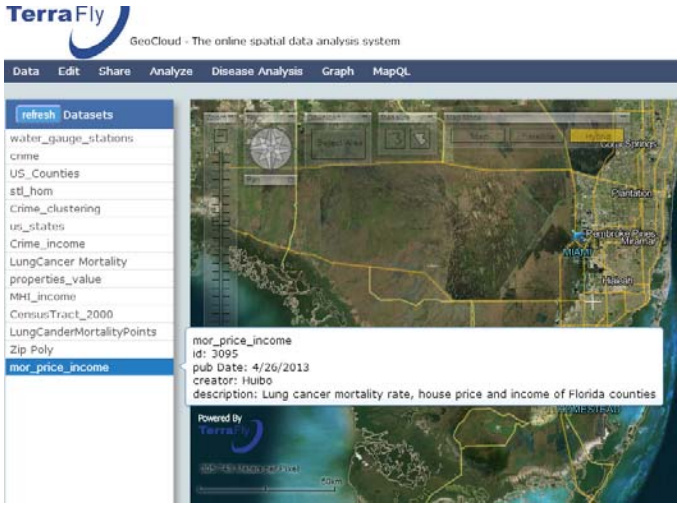

Fig. 7. Datasets in TerraFly GeoCloud

She can then choose the disease analysis button to draw a disease map. In this function, she can choose a legend group number; a disease map is displayed then, as shown in Figure 8.

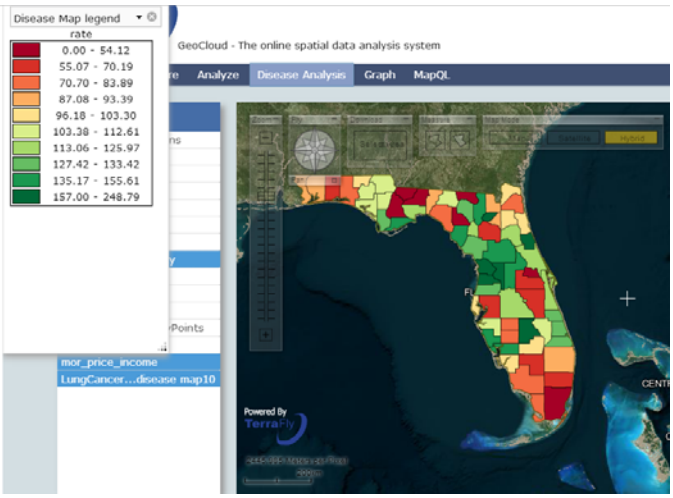

Fig. 8. Lung cancer disease map

From Figure 8 we see how this map, with legend at the top left corner, gives a direct summary of the disease data. For lung cancer in Florida, the mortality in the central region is higher and in the south is lower. However, the researcher cannot have an accurate analysis just from this one map. She can further choose the cluster and outliner function, which uses Local Moran's I to perform further analysis. This function provides three maps: local Moran's I map, z-value map, and p-value map. Figure 9 shows the p-value map, from which the researcher can know which counties form a statistically significant cluster and which counties are statistically significant outliners.

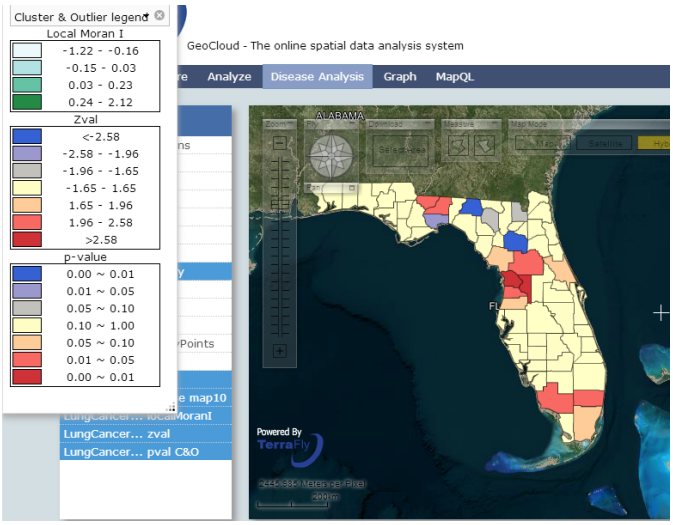

Fig. 9. P-value map of Local Moran I

Now the researcher may want to know what kind of relationship there is between lung cancer mortality and the median income of each county. For this purpose, she can use the median income dataset provided by the GeoCloud system, and apply to it the spatial auto-regression tool. Figure 10 shows the result of this model. From the result, we learn that when the mortality of surrounding areas increase by 1, the mortality of this county will increase of 0.233, and when the median income in the surrounding area increases by $10,000, the mortality of this county will decrease of 0.09.

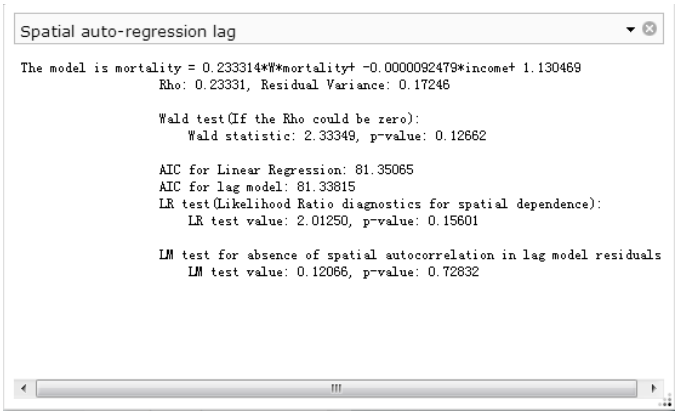

Fig. 10. Spatial auto-regression of lung cancer mortality and median income

VI. RELATED WORK AND PRODUCTS

Related software systems come in three categories: public health information providers, geo-statistic analytics, and survey analytics. A public health information provider dispenses disease maps and provides simple analysis functions. The dataset is fixed: users can neither upload their own datasets, nor perform complex analysis, such as cluster detection or regression.

ArcGIS and SaTScan are geo-statistic analytics software. ArcGIS spatial analyst does not provide an online service, nor does it aim at spatial epidemiology. SaTScan is used for spatial epidemiology, but the result is text only, and cannot be displayed as a map.

EPI Info is a well-respected survey analytics software that provides very good data analysis based on surveys. While featuring simple data visualization, it only provides limited geographic data analysis functions; it does not provide any online service.

The system is interfaced with our Health Informatics projects [16-23].

TABLE I. RELATED SPATIAL EPIDEMIOLOGY PRODUCTS

Name	Website	Description	Type
Georgia OASIS	http://oasis.state.ga.us/oasis/	Dataset collection, user can search data such as mortality and some behavioral surveys. This data can be displayed with map chart.	Public health information provider
Birtha	http://www.ehdp.com/birtha/	1 analyzing birth data; 2 calculate age-adjusted rate; 3 obtain rate by gender-age-race.	
ArcGIS	http://www.esri.com	This software provides multiple spatial	Geo statistic analyst

		analysis functions.	
SaTScan	http://www.satscan.org/	Provides statistic scan method for disease clusters.	
GeoViz Toolkit	http://www.geovista.psu.edu/geoviztoolkit/index.html	Provides disease visualization and Moran's I method .	
SpatialEpidemiology.net	http://www.spatialepidemiology.net/	Users can draw disease maps online.	
EPI Info	http://wwwn.cdc.gov/epiinfo/7/index.htm	Provides good data analysis functions based on surveys.	Survey analyst

VII. CONCLUSION AND FUTURE WORK

Our geospatial epidemiology system provides an important contribution to medicine and public health. It combines geospatial epidemiology and GIS technology, delivering analytics and user-friendly data visualization. This system provides comprehensive analysis methods, which are simple to use without any programming knowledge for spatial epidemiology. It also provides background datasets that may be of interest for epidemiologists. Our system leverages the map API provided by TerraFly GeoCloud in order to display the results of analysis; the distributed storage system allows users to upload their datasets, perform complex calculations online, and download the results.

Our future work includes implementing additional spatial methods for epidemiology, such as a spatial temporal cluster tool and point maps.

ACKNOWLEDGMENT

This material is based in part upon work supported by the National Science Foundation under Grant Nos. MRI CNS-0821345, MRI CNS-1126619, CREST HRD-0833093, I/UCRC IIP-1338922, I/UCRC IIP-0829576, RAPID CNS-1057661, RAPID IIS-1052625, MRI CNS-0959985, AIR IIP-1237818, SBIR IIP-1330943, FRP IIP-1230661, III-Large IIS-1213026, SBIR IIP-1058428, SBIR IIP-1026265, SBIR IIP-1058606, SBIR IIP-1127251, SBIR IIP-1127412, SBIR IIP-1118610, SBIR IIP-1230265, SBIR IIP-1256641, PIRE OISE-0730065, HECURA CCF-0938045, CAREER CNS-1253944, CAREER CNS-0747038, CSR CNS-1018262, HECURA CCF-0937964, I/UCRC IIP-0934364. Includes material licensed by TerraFly (http://terrafly.com) and the NSF CAKE Center (http://cake.fiu.edu).

REFERENCES

[1] P Elliott, Wartenberg D. Spatial epidemiology: current approaches and future challenges. Environ Health Perspect.2004;112:998–1006.

[2] Y. Lu, M. Zhang, T. Li, Y. Guang, N. Rishe, "Online Spatial Data Analysis and Visualization System"

[3] Armitage, Peter, Geoffrey Berry, and John Nigel Scott Matthews. Statistical methods in medical research. Wiley, 2008.

[4] Naphtali, et al. "System architecture for 3D terrafly online GIS." Multimedia Software Engineering, 2004. Proceedings. IEEE Sixth International Symposium on. IEEE, 2004.

[5] Lai, Alvin CK, Tracy L. Thatcher, and William W. Nazaroff. "Inhalation transfer factors for air pollution health risk assessment." Journal of the Air & Waste Management Association 50.9 (2000): 1688-1699.

[6] Wakefield, Jon, and Paul Elliott. "Issues in the statistical analysis of small area health data." Statistics in Medicine 18.17 - 18 (1999): 2377-2399.

[7] Kulldorff M. A spatial scan statistic. Communications in Statistics-Theory and methods. 1997;26:1481–96.

[8] Besag, Julian, and James Newell. "The detection of clusters in rare diseases."Journal of the Royal Statistical Society. Series A (Statistics in Society) (1991): 143-155.

[9] Kulldorff M, Nagarwalla N: Spatial disease clusters: detection and inference. *Statistics in Medicine* 1995, 14:799-810.

[10] Mantel, Nathan. "The detection of disease clustering and a generalized regression approach." Cancer research 27.2 Part 1 (1967): 209-220.

[11] Openshaw, Stan, et al. "A mark 1 geographical analysis machine for the automated analysis of point data sets." International Journal of Geographical Information System 1.4 (1987): 335-358.

[12] Getis, Arthur, and J. Keith Ord. "The analysis of spatial association by use of distance statistics." Geographical analysis 24.3 (1992): 189-206.

[13] Dubin, Robin, R. Kelley Pace, and Thomas G. Thibodeau. "Spatial autoregression techniques for real estate data." Journal of Real Estate Literature 7.1 (1999): 79-96.

[14] Kelejian, Harry H., and Ingmar R. Prucha. "A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances." The Journal of Real Estate Finance and Economics 17.1 (1998): 99-121.

[15] Moran, Patrick AP. "Notes on continuous stochastic phenomena." Biometrika 37.1/2 (1950): 17-23.

[16] Sagit Zolotov, Dafna Ben Yosef, Naphtali Rishe, Yelena Yesha, Eddy Karnieli. "Metabolic profiling in personalized medicine: bridging the gap between knowledge and clinical practice in Type 2 diabetes" Personalized Medicine, Vol. 8, No. 4, July 2011, pp. 445-456.

[17] 89. Naphtali Rishe, Carlos Espinal, Tajana Lucic, Yelena Yesha, Yaacov Yesha Kalai Mathee, Aileen Marty. "Geospatial Data for Intelligent Solutions in Public Health." e-Proceedings of Vaccinology 2012, Rio de Janeiro, September 3-7, 2012.

[18] Naphtali Rishe, Yelena Yesha, Yaacov Yesha, Tajana Lucic. "Intelligent solutions in public health: models and opportunities." Proceedings of the Second Annual International Conference on Tropical Medicine: Intelligent Solutions for Emerging Diseases. February 23-24, 2012, Miami, Florida.

[19] Naphtali Rishe, Yelena Yesha, Tajana Lucic. "Data Mining and Querying in Electronic Health Records." Proceedings of Up Close and Personalized, International Congress on Personalized Medicine (UPCP 2012), Florence, Italy, February 2-5, 2012.

[20] Yelena Yesha, Naphtali Rishe, Tajana Lucic. "Clinical-Genomic Analysis using Machine Learning Techniques to Predict Risk of Disease." Proceedings of Up Close and Personalized, International Congress on Personalized Medicine (UPCP 2012), Florence, Italy, February 2-5, 2012.

[21] Aniket Bochare, Aryya Gangopadhyay, Yelena Yesha, Anupam Joshi, Yaacov Yesha, Michael A. Grasso, Mary Brady, Naphtali Rishe. "Integrating Domain Knowledge in Supervised Machine Learning to Assess the Risk of Breast Cancer". International J of Medical Engineering and Informatics, 2013

[22] Rohit Kugaonkar, Aryya Gangopadhyay, Yelena Yesha, Anupam Joshi, Yaacov Yesha, Michael Grasso, Mary Brady and Naphtali Rishe. "Finding associations among SNPs for prostate cancer using collaborative filtering". DTMBIO-12: Proceedings of the ACM sixth international workshop on Data and text mining in biomedical informatics. Hawaii, USA. October 29, 2012. ACM New York, NY.

[23] Ron Ribitzky, Yelena Yesha, Eddy Karnieli, Naphtali Rishe. "Knowledge Mining & Bio-informatics Techniques to Advance Personalized Diagnostics & Therapeutics". Report to the U.S. National Science Foundation (NSF) on the Outcomes and Consensus Recommendations of the NSF-sponsored International Workshop, February 2012, in Florence, Italy, http://CAKE.fiu.edu/HIT-papers/Book_post_NSF_Workshop_Knowledge_Mining_and_Bioinformatics_Techniques_to_Advance_Personalized_Diagnostics_and_Therapeutics.pdf