Online Spatial Data Analysis and Visualization System

Mingjin Zhang, Florida International University
Huibo Wang, Florida International University
Yun Lu, Florida International University
Tao Li, Florida International University
Yudong Guang, Florida International University
Chang Liu, Florida International University
Erik Edrosa, Florida International University
Naphtali Rishe, Florida International University

With the exponential growth of the usage of web map services, the geo data analysis has become more and more popular. This paper develops an online spatial data analysis and visualization system, TerraFly GeoCloud, which facilitates end users to visualize and analyze spatial data, and to share the analysis results. Built on the TerraFly Geo spatial database, TerraFly GeoCloud is an extra layer running upon the TerraFly map and can efficiently support many different visualization functions and spatial data analysis models. Furthermore, users can create unique URLs to visualize and share the analysis results. TerraFly GeoCloud also enables the MapQL technology to customize map visualization using SQL-like statements. The system is available at http://terrafly.fiu.edu/GeoCloud/.

Categories and Subject Descriptors: C.2.2 [**Computer-Communication Networks**]: Network Protocols

General Terms: Design, Algorithms, Performance

Additional Key Words and Phrases: Geospatial analysis, GIS, Visualization, Big Data

ACM Reference Format:
Mingjin Zhang, Huibo Wang, Yun Lu, Tao Li, Yudong Guang, Chang Liu, Erik Edrosa, Naphtali Rishe, 2010. Online Spatial Data Analysis and Visualization System. *ACM Trans. Embedd. Comput. Syst.* 9, 4, Article 39 (March 2010), 25 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

With the exponential growth of the World Wide Web, there are many domains, such as water management, crime mapping, disease analysis, and real estate, open to Geographic Information System (GIS) applications. The Web can provide a giant amount of information to a multitude of users, making GIS available to a wider range of public users than ever before. Web-based map services are the most important application of modern GIS systems. For example, Google Maps currently has more than 350 mil-

This material is based in part upon work supported by the National Science Foundation under Grant Nos. I/UCRC IIP-1338922, AIR IIP-1237818, SBIR IIP-1330943, III-Large IIS-1213026, MRI CNS-0821345, MRI CNS-1126619, CREST HRD-0833093, I/UCRC IIP-0829576, MRI CNS-0959985, FRP IIP-1230661, SBIR IIP-1058428, SBIR IIP-1026265, SBIR IIP-1058606, SBIR IIP-1127251, SBIR IIP-1127412, SBIR IIP-1118610, SBIR IIP-1230265, SBIR IIP-1256641. Includes material licensed by TerraFly (http://terrafly.com) and the NSF CAKE Center (http://cake.fiu.edu).
Author's addresses: M. Zhang, H. Wang, Y. Lu, T. Li, Y. Guang, E. Edrosa, N. Rishe, Computer Information Science Department, Florida International University;

DOI: http://dx.doi.org/10.1145/0000000.0000000

lion users. There are also a rapidly growing number of geo-enabled applications which utilize web map services on traditional computing platforms as well as the emerging mobile devices.

However, due to the highly complex and dynamic nature of GIS systems, it is quite challenging for the end users to quickly understand and analyze the spatial data, and to efficiently share their own data and analysis results to others. First, typical geographic visualization tools are complicated and fussy with a lot of low-level details, thus they are difficult to use for spatial data analysis. Second, the analysis of large amount spatial data is very resource-consuming. Third, current spatial data visualization tools are not well integrated for map developers and it is difficult for end users to create the map applications on their own spatial datasets.

To address the above challenges, this paper presents TerraFly GeoCloud, an online spatial data analysis and visualization system, which allows end users to easily visualize and share various types of spatial data.

— First, TerraFly GeoCloud can accurately visualize and manip-ulate point and polygon spatial data with just a few clicks.
— Second, TerraFly GeoCloud em-ploys an analysis engine to support the online analysis of spatial data, and the visualization of the analysis results. Many different spatial analysis functionalities are provided by the analysis engine.
— Third, based on the TerraFly map API, TerraFly GeoCloud offers a MapQL language with SQL-like statements to execute spatial queries, and render maps to visualize the customized query results.

Our TerraFly GeoCloud online spatial data analysis and visualization system is built upon the TerraFly system using TerraFly Maps API and JavaScript TerraFly API addons in a high performance cloud Environment. The function modules in the analysis engine are im-plemented using C and R language and python scripts. Comparing with current GIS applications, our system is more user-friendly and offers better usability in the analysis and visualization of spatial data. The system is available at http://terrafly.fiu.edu/GeoCloud/.

The previous paper[Lu et al. 2013] focused on visualization solution such as the map rendering and visualizing spatial data. After the previous paper, we added many spatial analysis functions and made the result visualization more interactive. With these changes Geocloud became more intelligent and can be applied on many domains, such as disease analysis, crime analysis and real estate analysis.

The rest of this paper is organized as follows: Section 2 presents the background and motivation; Sections 3 describes the architecture of TerraFly GeoCloud; Section 4 describes the visualization and analysis methods in TerraFly GeoCloud; Section 5 presents case study on the online spatial analysis; Section 6 discusses the related work; and finally Section 7 concludes the paper.

2. BACKGROUND

TerraFly, a high-performance web-based geographic information system, is developed by High-Performance Database Research Center (HPDRC) in Florida International University. TerraFly provides multiple functions related to geospatial such as providing customized aerial photography, satellite photography, topographic maps, data layers, data report and data analysis functions. User can use these functions by inputting street, city and state, etc.

TerraFly as an internet visualization geographic information system has 40 TB database of aerial imagery and spatial data and robust spatial cloud-computing environment. TerraFly also provides multiple GIS-oriented Application Programming Interface (API).

TerraFly server contains 40TB high resolution data collection such as 1-meter aerial photography of United States and 3-inch to 1-foot full color recent imagery of major cities. The vector data collection which includes 400 million geo-located objects, 50 billion data fields, 40 million polylines, and 120 million polygons covers all United State and Canada roads. It also contains some big datasets such as the US Census demographic and socioeconomic datasets, parcels datasets about property lines and ownership data, business datasets about company stats and management roles and contacts, physician datasets and hundreds of other data sets. Its cloud-computing environment can provide strong computing and storage ability.

TerraFly map API has been using in several products such as water management, realtor, geo-photos, weather, airline, autopilot and travel. TerraFly API can be easily deployed on web service. User can visualize map through map API on web browser and without installing any software. TerraFly map API provides query and visualizing interface which are easy to use. It allows user to add layers to customize visualization. [Rishe et al. 2001][Rishe et al. 2005]

3. TERRAFLY GEOCLOUD

Figure 1 shows the system architecture of TerraFly GeoCloud. Based on the current TerraFly system including the Map API and all sorts of TerraFly data, we developed the TerraFly GeoCloud system to perform online spatial data analysis and visualization. In TerraFly GeoCloud, users can import and visualize various types of spatial data (data with geo-location information) on the TerraFly map, edit the data, perform spatial data analysis, and visualize and share the analysis results to others. Available spatial data sources in TerraFly GeoCloud include but not limited to demographic census, real estate, disaster, hydrology, retail, crime, and disease. In addition, the system supports MapQL, which is a technology to customize map visualization using SQL-like statements.

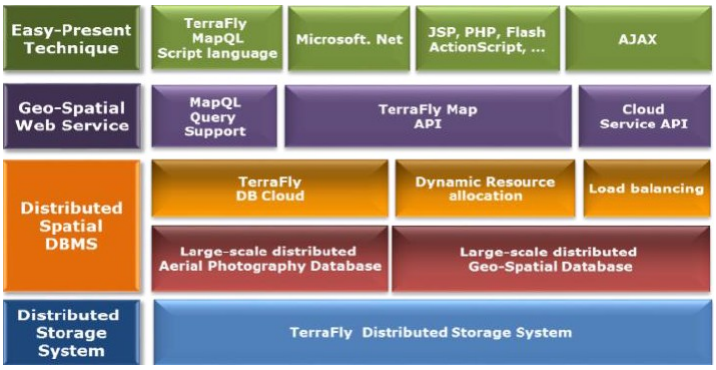


Fig. 1: The Architecture pf TerraFly GeoCloud

The spatial data analysis functions provided by TerraFly GeoCloud include spatial data visualization (visualizing the spatial data), spatial dependency and autocorrelation (checking for spatial dependencies), spatial clustering (grouping similar spatial objects),spatial regression, measuring Geographic Distribution and Kriging (geostatistical estimator for unobserved locations).

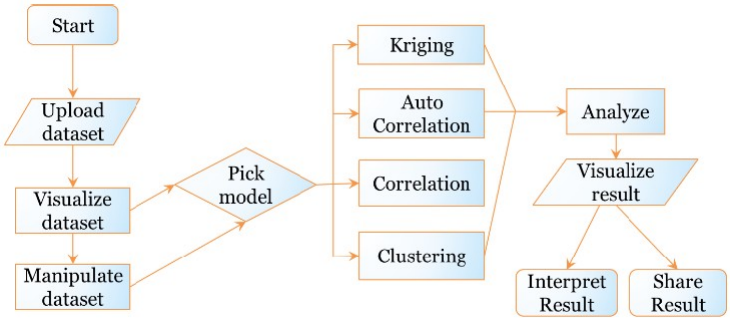


Fig. 2: The Workflow of TerraFly Geocloud

Figure2 shows the data analysis workflow of the TerraFly GeoCloud system. Users first upload datasets to the system, or view the available datasets in the system. They can then visualize the data sets with customized appearances. By Manipulate dataset, users can edit the dataset and perform pre-processing (e.g., adding more columns). Followed by pre-processing, users can choose proper spatial analysis functions and perform the analysis. After the analysis, they can visualize the results and are also able to share them with others.

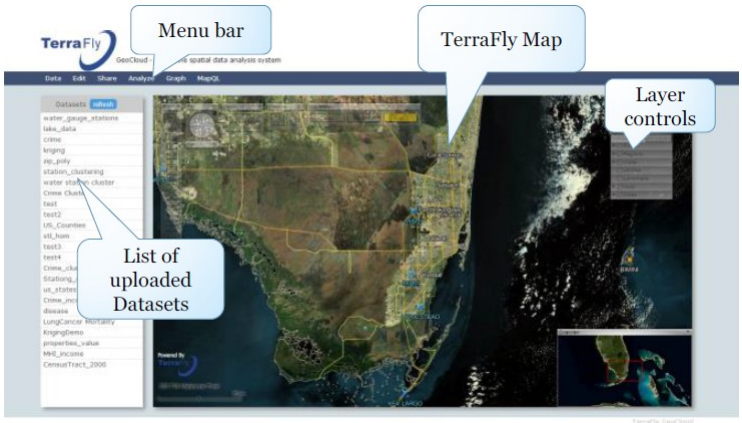


Fig. 3: Interface of TerraFly Geocloud

Figure 3 showed the interface of the TerraFly GeoCloud system. The top bar is the menu of all functions, including Data, analysis, Graph, Share, and MapQL. The left side shows the available datasets, including both the uploaded datasets from the user and the existing datasets in the system. The right map is the main map from TerraFly. This map is composed by TerraFly API, and it includes a detailed base map and diverse overlays which can present different kinds of geographical data.

TerraFly GeoCloud also provides MapQL spatial query and render tools. MapQL supports SQL-like statements to realize the spatial query, and after that, render the map according to users inputs. MapQL tools can help users visualize their own data

using a simple statement. This provides users with a better mechanism to easily visualize geographical data and analysis results.

4. VISUALIZATION AND ANALYSIS

TerraFly GeoCloud integrates spatial data mining and data visualization. The spatial data mining results can be easily visualized. In addition, visualization can often be incorporated into the spatial mining process.

4.1. Spatial Data Visualization

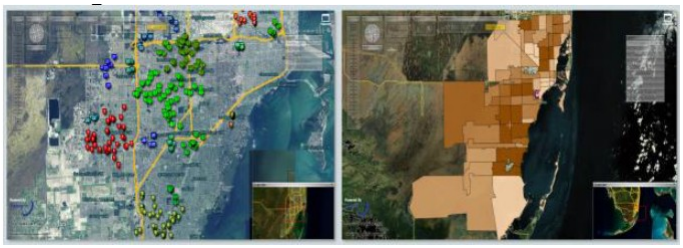


Fig. 4: Spatial Data Visualization: Point data and Polygon Data

For spatial data visualization, the system supports both point data and polygon data and users can choose color or color range of data for displaying. As shown in Figure 4, the point data is displayed on left, and the polygen data is displayed on the right. The data labels are shown on the bajse map as extra layers for point data, and the data polygons are shown on the base map for polygon data. Many different visualization choices are supported for both point data and polygon data. For point data, user can customize the icon style, icon color or color range, label value and so on. For polygon data, user can customize the fill color or color range, fill alpha, line color, line width, line alpha, label value and so on.

4.2. Spatial dependency and Auto-Correlation

Spatial dependency is the co-variation of properties within geographic space: characteristics at proximal locations that appear to be correlated, either positively or negatively. Spatial dependency leads to the spatial autocorrelation problem in statistics.[De Knegt et al. 2010] Spatial autocorrelation is more complex than one-dimensional autocorrelation because spatial correlation is multi-dimensional (i.e. 2 or 3 dimensions of space) and multi-directional. The TerraFly GeoCloud system provides auto-correlation analysis tools to discover spatial dependencies in a geographic space, including global and local clusters analysis where Moran's I measure is used .[Li et al. 2007] Formally, Morans I, the slope of the line, estimates the overall global degree of spatial autocorrelation as follows:

$$I = \frac{n}{\sum_i^n \sum_j^n w_{ij}} * \frac{\sum_i^n \sum_j^n w_{ij}(y_i - \hat{y})(y_i - \hat{y})}{\sum_i^n (y_i - \hat{y})^2}, \quad (1)$$

where w_{ij} is the weight, $w_{ij} = 1$ if locations i and j are adjacent and zero otherwise $w_{ii} = 0$ (a region is not adjacent to itself).y_i and $\hat{y}$ are the variable in the ith location and the mean of the variable, respectively. n is the total number of observations. Morans I is used to test hypotheses concerning the correlation, ranging between 1.0 and $+1.0$. Morans I measures can be displayed as a checkerboard where a positive Morans I measure indicates the clustering of similar values and a negative Morans I measure

indicate dissimilar values. TerraFly GeoCloud system provides auto-correlation analysis tools to check for spatial dependencies in a geographic space, including global and local clusters analysis

Fig. 5: Average properties price by zip code in Miami

Local Morans I is a local spatial autocorrelation statistic based on the Morans I statistic. It was developed by Anselin as a local indicator of spatial association or LISA statistic [Anselin 1995]. The fact that Moran's I is a summation of individual cross products is exploited by the "Local indicators of spatial association" (LISA) to evaluate the clustering in those individual units by calculating Local Moran's I for each spatial unit and evaluating the statistical significance for each I_i. From the previous equation we then obtain:

$$I_i = z_i \sum_{j}^{n} w_{ij} z_{ij}, \tag{2}$$

where zi are the deviations from the mean of yi, and the weights are row standardized. Figure 5 shows an example of spatial auto-correlation analysis on the average properties price by zip code data in Miami (polygondata). Each dot here in the scatterplot corresponds to one zip code. The first and third quadrants of the plot represent positive associations (high-high and low-low), while the second and fourth quadrants represent associations (low-high, high-low). For example, the green circle area is in the low-high quadrants. The density of the quadrants represents the dominating local spatial process. The properties in Miami Beach are more expensive, and are in the high-high area.

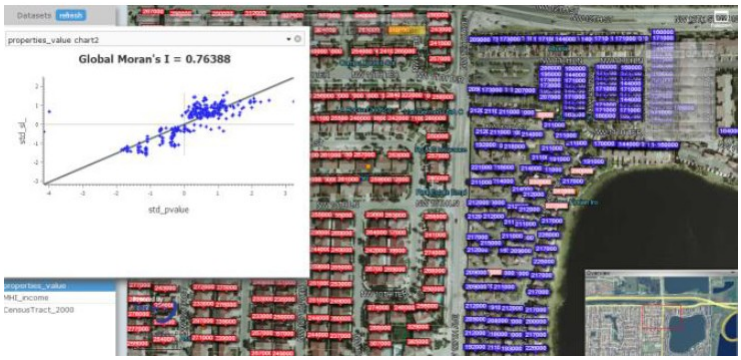


Fig. 6: Properties value in Miami

Figure 6 presents the auto-correlation analysis results on the individual properties price in Miami (point data). Each dot here in the scatterplot corresponds to one property. As the figure shows, the properties near the big lake are cheaper, while the properties along the west are more expensive.

4.3. Spatial Data Clustering

Spatial data clustering is trying to find out some objects gathering together with some similar features and all the elements in the same cluster must be close in geography.

DBSCAN. The TerraFly GeoCloud system supports the DBSCAN (for density-based spatial clustering of applications with noise) data clustering algorithm.[Ester et al. 1996] It is a density-based clustering algorithm because it finds a number of clusters starting from the estimated density distribution of corresponding nodes. DBSCAN requires two parameters as the input: eps and the minimum number of points required to form a cluster minPts. It starts with an arbitrary starting point that has not been visited so far. This point's neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as a noise point.[Ester et al. 1996] If a point is found to be a dense part of a cluster, its neighborhood is also part of that cluster. Hence, all points that are found within the neighborhood are added. This process continues until the density-connected cluster is completely identified. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise points.[Bilodeau et al. 2005] Figure 7 shows an example of DBSCAN clustering on the crime data in Miami. As shown in Figure 7, each point is an individual crime record marked on the place where the crime happened, and the number displayed in the label is the crime ID. By using the clustering algorithm, the crime records are grouped, and different clusters are represented by different colors on the map.

Fig. 7: DBSCAN clustering on the crime data in Miami

Cluster Detection. Kulldorff & Nagarwalla(KN)[Kulldorff 1997] provides a method to do cluster detection. KN method is implemented by scanning all the area using a circular zones of variable size. KN method is widely used in spatial epidemiology.

The steps of 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 circles and sort the LLR; 3. Get some large LLR then use Monte Carlo method to calculate P-value of them. The Log Likelihood Ratio can be calculated as follow:

$$LLR = max_j(\frac{Y_j}{E_j})^{Y_j}(\frac{Y_+ + Y_j}{Y_+ - E_j})^{Y_+ - Y_J} I(Yj > Ej), \tag{3}$$

Where Y_j denotes the observed number of instance in circle area, $Y+$ denotes the number of instance in all the area, E_j denotes the expected number of instance in circle area.

Figure 8a shows the result of lung cancer cluster map in Florida. The red points indicate the disease cluster where the unusual disease case happened. The number in the red point is the p-value of each area.[Elliott and Wartenberg 2004]

HotSpot. HotSpot analysis function using Gi* statistic method is to detect the hot cluster where has high value and clod cluster where has low value.

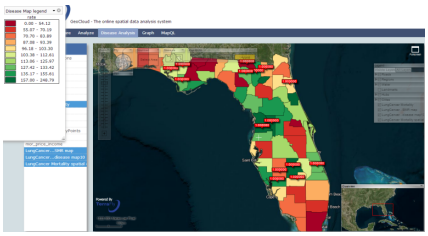
Figure 8b shows the result of hotspot cluster map of lung cancer mortality in Florida. From this map, we can know the center part which is covered by red is hot cluster with significant p value. And four counties in the south part consistent clod cluster with statistic significant p value.

Cluster and Outlier. Cluster and outliner analysis recognizes spatial cluster where all the point or polygon have similar attributes values. This methods can also recognize the outliner whoes attributes values are different from around.In this method, local moran's I map, z-value map and p-value map are provided.

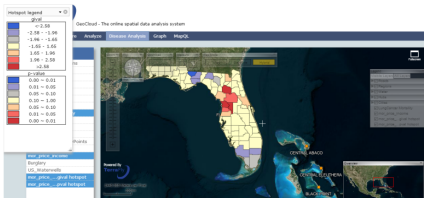
4.4. Spatial Regression

Regression tools can be used to estimate relationships between other datasets such as socioeconomic data.

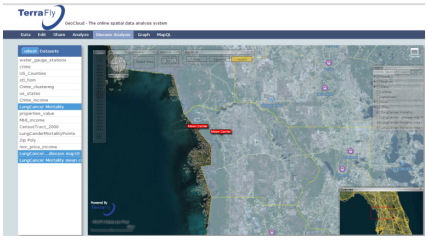
Linear Regression. TerraFly provides linear regression tool with multiple tests, such as global morans I test.



(a) KN cluster detection on lung cancer in Florida



(b) Hotspot clustering on lung cancer in Florida



(c) Center point and weighted center point

Fig. 8: Spatial Analysis in Geocloud

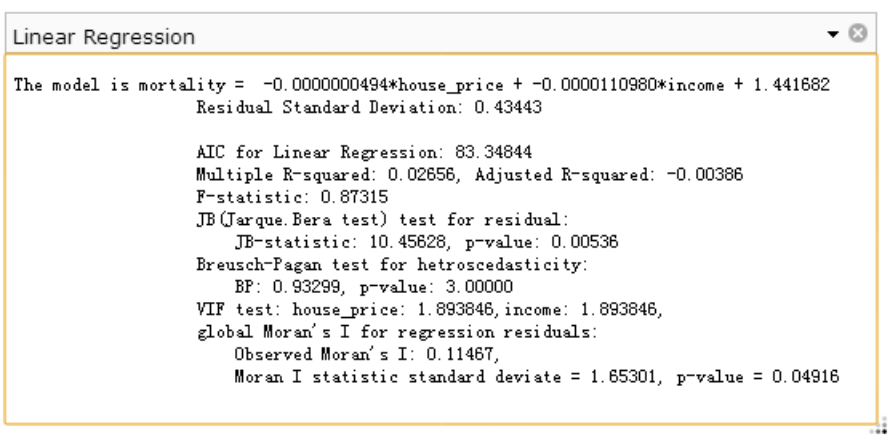


Fig. 9: Linear regression tool on lung cancer in Florida

Figure 9 shows linear regression result between mortality and median house price and median income. But global Morans I test indicates that the residual is geo-correlated and linear regression model is not fit for this problem.

Spatial auto-regression. In spatial auto-regression, a lag model and an error model are provided. The spatial auto-regression lag model can be calculated as follows:

$$Y = \rho W y + x\beta + \epsilon, \tag{4}$$

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

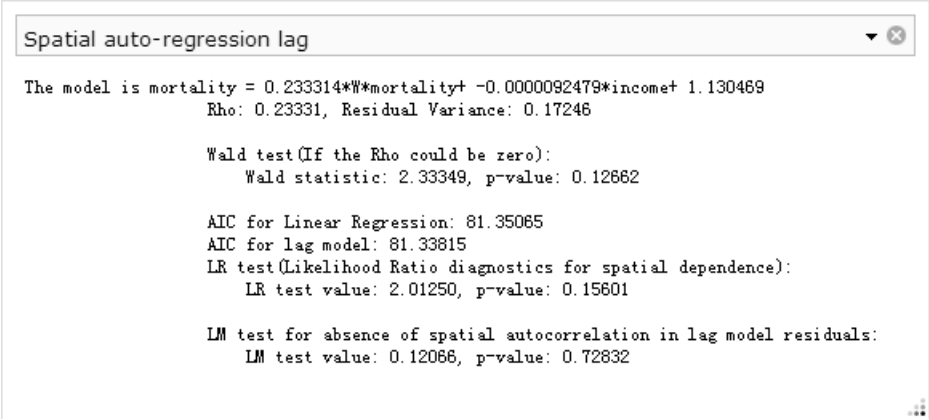


Fig. 10: Spatial auto-regression lag model on lung cancer in Florida

Figure 10 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.[Dubin et al. 1999][Kelejian and Prucha 1998]

4.5. 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:

$$X = \frac{\Sigma_i w_i x_i}{\Sigma_i w_i}, Y = \frac{\Sigma_i w_i y_i}{\Sigma_i w_i}, \tag{5}$$

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.

Figure 8c 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.

Besides the center/median point function, GeoCloud includes distributional trends and standard distance.

4.6. Spatial Interpolation Method

Kriging is a geo-statistical estimator that infers the value of a random field at an unobserved location (e.g. elevation as a function of geographic coordinates) from samples (see spatial analysis) [Stein 1999]

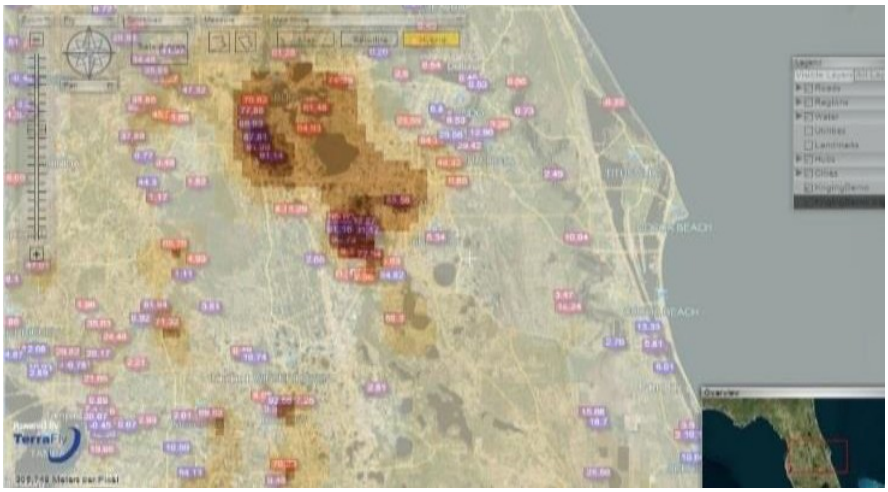


Fig. 11: Kriging data of the water level in Florida

Figure 11 shows an example of Kriging. The data set is the water level from water stations in central Florida. Note that not all the water surfaces are measured by water stations. The Kriging results are estimates of the water levels and are shown by the yellow layer.

5. CUSTOMIZED MAP VISUALIZATION

TerraFly GeoCloud also provides MapQL spatial query and render tools, which supports SQL-like statements to facilitate the spatial query and more importantly, render the map according users requests. This is a better interface than API to facilitate developer and end user to use the TerraFly map as their wish. By using MapQL tools, users can easily create their own maps.

5.1. Implementation

The implementation of MapQL is shown in Figure 12. The input of the whole procedure is MapQL statements, and the output is map visualization rendered by the MapQL engine.

Shown in Figure 12, the first step is syntax check of the statements. Syntax check guarantees that the syntax conforms to the standard, such as the spelling-check of the reserved words. Semantic check ensures that the data source name and metadata which MapQL statements want to visit are correct. After the above two checks, system will parse the statements and store the parse results including the style information into a spatial database. The style information includes where to render and what to render. After all the style information is stored, system will create style configuration objects for render. The last step is for each object, load the style information form spatial database and render to the map according to the style information.

We implemented the MapQL tools using C++. For the last step which is rendering the objects to the map visualization, we employed the TerraFly map render engine.[Wang 2011]

For example, if we want to query the house prices near Florida International University, we use MapQL statements like figure 13

There are four reseverd words in the statements, T_ICON_PATH , T_LABEL, T_LABEL_SIZE , and GEO. We use T_ICON_PATH to store the customized icon. Here

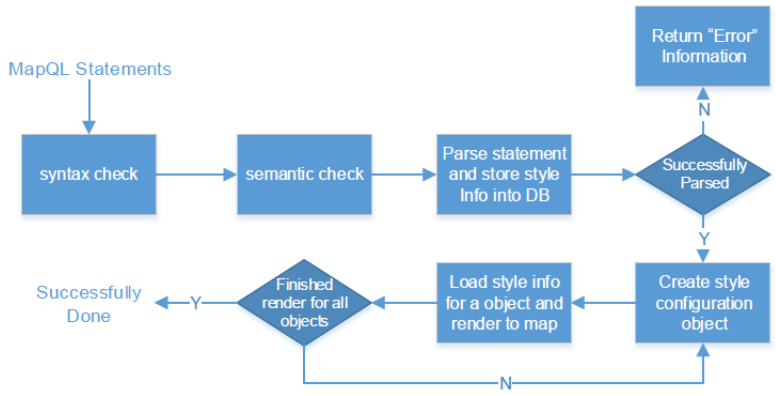


Fig. 12: MapQL implementation

```
SELECT
    '/var/www/cgi-bin/house.png' AS T_ICON_PATH,
r.price AS T LABEL,
    '15' AS T LABEL SIZE,
r.geo AS GEO
FROM
realtor 20121116 r
WHERE
ST Distance(r.geo,  GeomFromText('POINT(-80.376283 25.757228)')) <
0.03;
```

Fig. 13: Query house prices using MapQL

we choose a local png file as icon. T_LABEL denotes that icon label that will be shown on the map, . T_LABEL_SIZE is the pixel size of the label; and GEO is the spatial search geometry.

The statement goes through the syntax check first. If there is incorrect usage of reserved words or wrong spelling of the syntax, it will be corrected or Error information will be sent to users. For example, if the spelling of select is not correct, Error information will be sent to user. Semantic check makes sure that the data source name realtor_20121116 and metadata r. price and r.geo are exist and available.

After the checks, the system parsed the statements. The SQL part will return corresponding results including the locations and names of nearby objects, the MapQL part will collect the style information like icon path and icon label style. Both of them are stored into a spatial database. The system then created style configuration objects for query results. The last step is rendering all the objects on the map visualizations. The style information needed includes icon picture and label size, and the data information includes label value and location (Lat, Long).

Figure 14 shows the result of this query. Please be noticed that the unit of the distance func-tion in all the demos is Lat-Long.

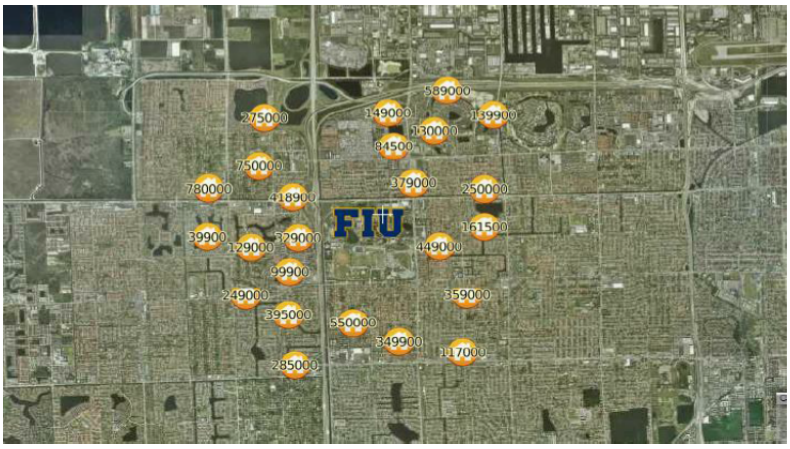


Fig. 14: Result of query house prices using MapQL

5.2. Other Examples

Figure 15shows all the hotels along a certain street within a certain distance and also displays the different stars of the hotels. The MapQL statement for this query is listed below:

Figure 16shows the traffic of Santiago where the colder the color is, the faster the traffic is, the warmer the color is, and the worse the traffic is. The MapQL statement is listed below:

Figure 17 shows the different average incomes with in different zip codes. In this demo, users can customize the color and style of the map layers, different color stand for different average incomes. And the MapQL statement is listed below:

All these examples demonstrate that in TerraFly GeoCloud, users can easily create different map applications using simple SQL-like statements.

6. CASE STUDY

In this section, we present some case study on using TerraFly GeoCloud for spatial data analysis and visualization. We using two kinds of data set, one is Florida property data, the other is Florida Lung cancer mortality to show how to apply Geocloud analysis and visualization function on several domains.

6.1. Florida property analysis

As discussed in 4.2.1, we know the results of auto correlation can be shown in a scatter diagram, where the first and third quadrants of the plot represent positive associations, while the second and fourth quadrants represent negative associations. The second quadrant stands for low-high which means the value of the object is low and the values of surrounding objects are high. A lay user whose name is Erik who has some knowledge about the database and data analysis wanted to invest a house property in Miami with a good appreciation potential. By using TerraFly GeoCloud, he may obtain some ideas about where to buy. He believes that if a property itself has low price and the surrounding properties have higher values, then the property may have good appreciation potential, and is a good choice for investment. He wants to first identify such properties and then do a field trip with his friends and the realtor agent.

```
SELECT
    CASE
        WHEN star >= 1 and star < 2 THEN '/var/www/cgi-bin/hotel_1star.png'
        WHEN star >= 2 and star < 3 THEN '/var/www/cgi-bin/hotel_2stars.png'
        WHEN star >= 3 and star < 4 THEN '/var/www/cgi-bin/hotel_3stars.png'
        WHEN star >= 4 and star < 5 THEN '/var/www/cgi-bin/hotel_2stars.png'
        WHEN star >= 5 THEN '/var/www/cgi-bin/hotel_2stars.png'
        ELSE '/var/www/cgi-bin/hotel_0star.png'
    END AS T_ICON_PATH,
h.geo AS GEO
FROM
osm_fl o
LEFT JOIN
hotel_all h
ON
ST_Distance(o.geo, h.geo) < 0.05
WHERE
    o.name = 'Florida Turnpike';
```

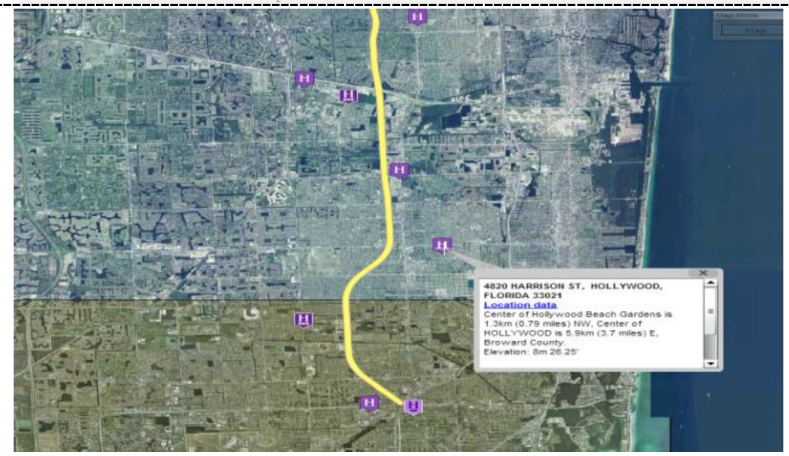


Fig. 15: Query hotel data along the line

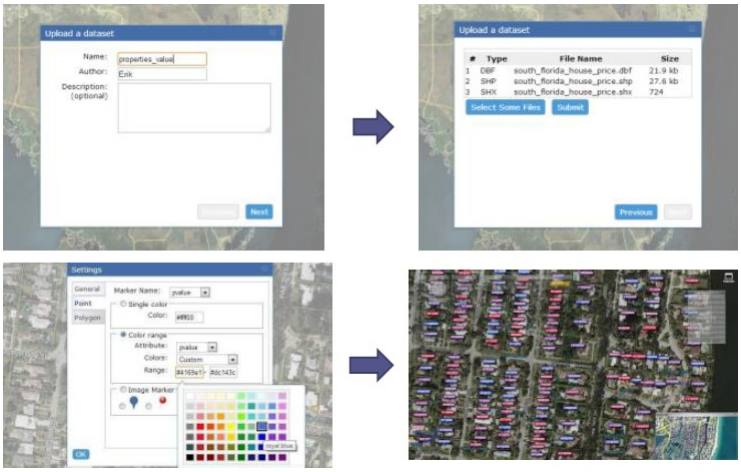


Fig. 18: Data Set Upload and Visualization

To perform the task, first, Erik checked the average property prices by zip code in Miami which is shown in Figure 6. He found the green circled area in the low-high quad-

```
SELECT
    CASE
        WHEN speed >= 50 THEN 'color(155, 188, 255)'
        WHEN speed >= 40 and speed < 50 THEN 'color(233, 236, 255)'
        WHEN speed >= 30 and speed < 40 THEN 'color(255, 225, 198)'
        WHEN speed >= 20 and speed < 30 THEN 'color(255, 189, 111)'
        WHEN speed >= 10 and speed < 20 THEN 'color(255, 146, 29)'
        WHEN speed >= 5 and speed < 10 THEN 'color(255, 69, 0)'
        WHEN speed >= 0 and speed < 5 THEN 'color("red")'
else 'color("grey")'
    END AS T_FILLED_COLOR,
    '3' AS T_THICKNESS,
GEO
FROM santiago_traffic;
```

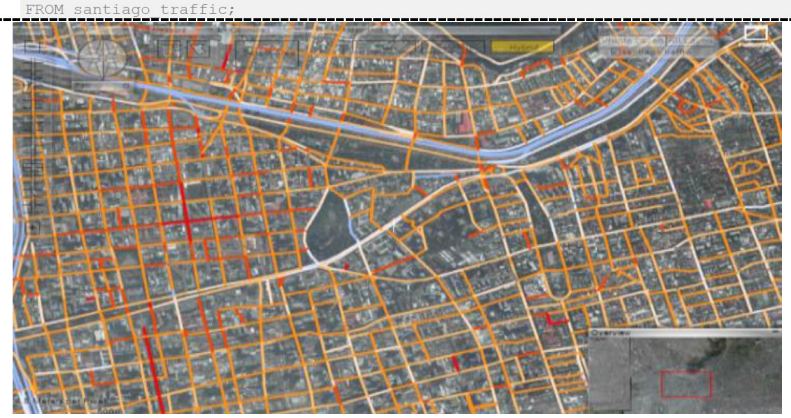


Fig. 16: Query traffic data of Santiago

rants, which means that the average price of properties of this area is lower than the surrounding areas. Then, Erik wanted to obtain more insights on the property price in this area. He uploaded a detailed spatial data set named as south_florida_house_price into the TerraFly GeoCloud system as shown in Figure 18. He customized the label color range as the properties price changes. And then, he chose different areas in the green circled area in Figure 19 to perform the auto-correlation analysis.

Fig. 19: Properties in Miami

Finally, he found an area shown in Figure 20, where there are some good properties in the low-high quadrants (in yellow circles) with good locations. And one interesting observation is, lots of properties along the road Gratigny Pkwy has lower prices.

```
SELECT
u.geo AS GEO,
u.zip AS T_LABEL,
    '0.7' AS T_OPACITY,
    '15' AS T_LABEL_SIZE,
'color("blue")' AS T_BORDER_COLOR,
    CASE
        WHEN avg(i.income) < 30000 THEN 'color(155, 188, 255)'
        WHEN avg(i.income) >= 30000 and avg(i.income) < 50000 THEN 'color(233,
236, 255)'
        WHEN avg(i.income) >= 50000 and avg(i.income) < 70000 THEN 'color(255,
225, 198)'
        WHEN avg(i.income) >= 70000 and avg(i.income) < 90000 THEN 'color(255,
189, 111)'
        WHEN avg(i.income) >= 90000 and avg(i.income) < 110000 THEN 'color(255,
146, 29)'
        WHEN avg(i.income) >= 110000 and avg(i.income) < 130000 THEN 'color(255,
69, 0)'
        WHEN avg(i.income) >= 130000 THEN 'color("red")'
else 'color("grey")'
    END AS T_FILLED_COLOR
FROM
us_zip u left join income i
ON
ST_Within(i.geo, u.geo)='t'
GROUP BY
u.geo, u.zip;
```

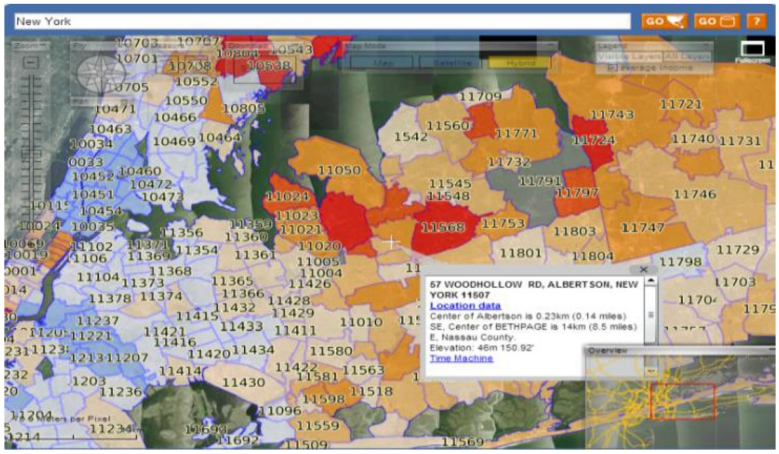


Fig. 17: Query average incomes

He was then very excited and wanted to do a query to find all the cheap properties with good appreciation potential along the Gratigny Pkwy. Erik composed the MapQL statements like:

```
SELECT
    CASE
        WHEN h.pvalue >= 400000 THEN '/var/www/cgi-bin/redhouse.png'
        WHEN h.pvalue >= 200000 and h.pvalue < 400000 THEN '/var/www/cgi-
bin/bluehouse.png'
        WHEN h.pvalue >= 100000 and h.pvalue < 200000 THEN '/var/www/cgi-
bin/greenhouse.png'

        ELSE '/var/www/cgi-bin/darkhouse.png'
    END AS T_ICON_PATH,
h.geo AS GEO
FROM
osm_fl o
LEFT JOIN
south florida house price h
ON
ST_Distance(o.geo, h.geo) < 0.05
WHERE
    o.name = 'Gratigny Pkwy' AND
    h.std_pvalue<0 AND
    h.std_sl_pvalue>0;
```

Fig. 20: MapQL statement

Fig. 21: MapQL results

The Figure 21 presents the final results of the MapQL statements. Finally, Erik sent the URL of the map visualization out by email, waiting for the response of his friends and the realtor agent.

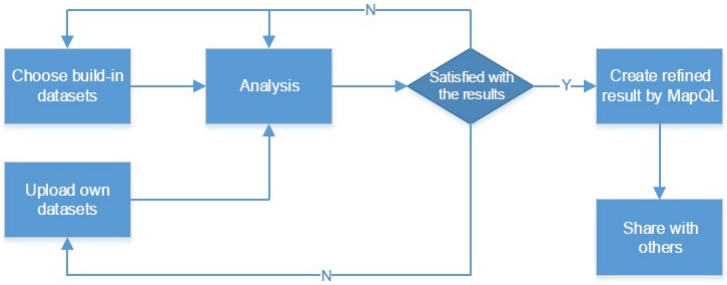


Fig. 22: The flow path of Erik case

Figure 22 illustrates the whole workflow of the case study. In summary, Erik first viewed the system build-in datasets, conducted the data analysis, and then he identified properties of interest. He then composed MapQL statements to create his own map visualization to share with his friends. The case study demonstrates that TerraFly GeoCloud supports the integration of spatial data analysis and visualization and also offers user-friendly mechanisms for customized map visualization.

6.2. Florida Lung Cancer analysis

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 23.

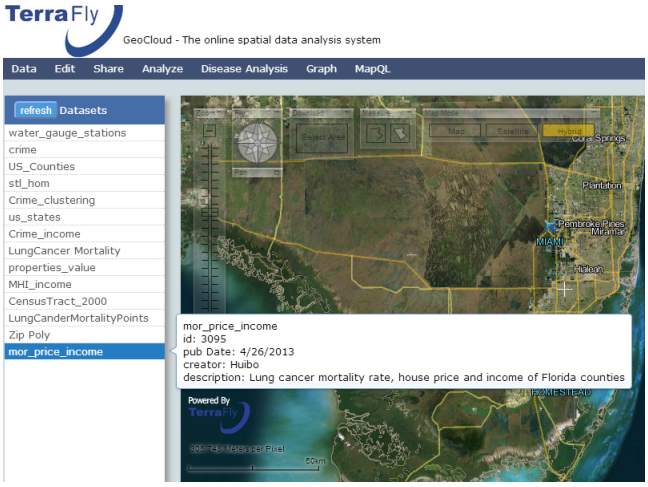


Fig. 23: Datasets in TerraFly GeoCloud

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 24.

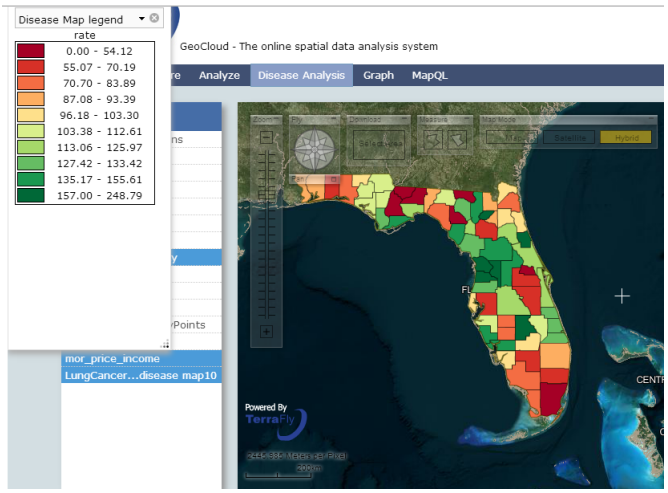


Fig. 24: Lung Cancer disease map

From Figure 24 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 Morans I to perform further analysis. This function provides three maps: local Morans I map, z-value map, and p-value map. Figure 25 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.

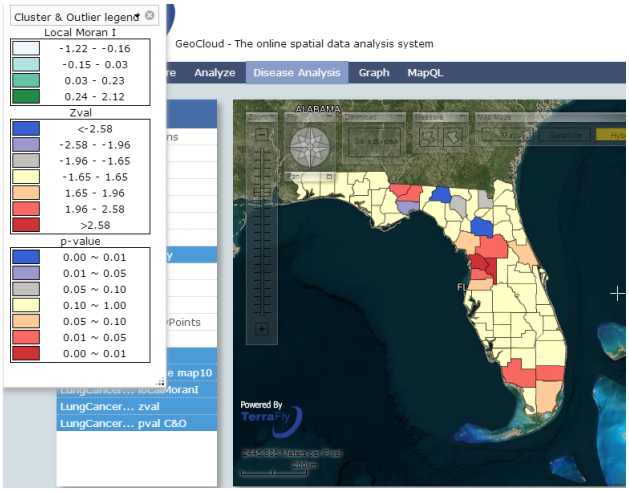


Fig. 25: P-value map of Local Moran I

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 26 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.

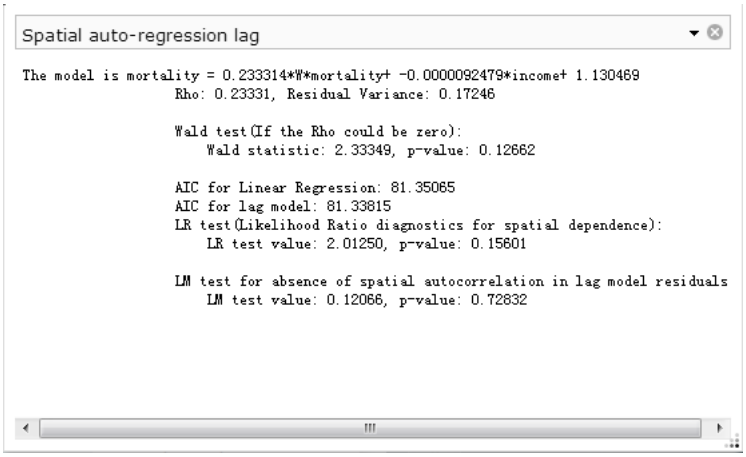


Fig. 26: Spatial auto-regression of lung cancer mortality and median income

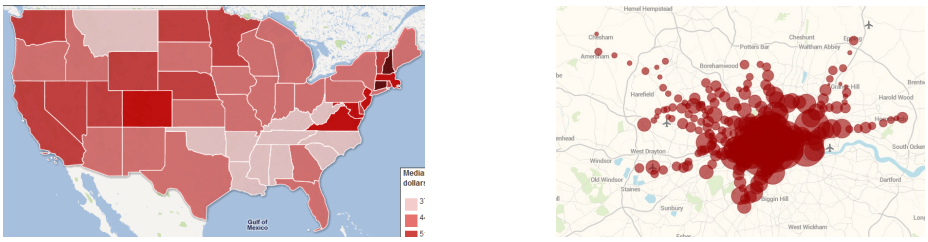
7. RELATED WORK AND PRODUCTS

7.1. Spatial Data Visualization

Data visualization is the study to show the abstract data as image and make it easy to be summarized by human beings.[Old 2002] The abstract data contains numerical and text data. The visualization is interactivce between human beings and computer. Data visualization is applied in many domains such as data analysis, data mining and economics. There are a lot of effective data visualization approaches which contain histogram, pie chart and scatter plots, etc. These approaches make the data easy to be interpreted and analyzed.[Spence and Press 2000]

Spatial data visualization as one kind of data visualization is to visualize the geo-related data. Map is the first spatial data visualization method, but as computer technology developing, user has more and more spatial data, such as demographic data related to some place, need to be demonstrated. Then map became a background for user own spatial data visualizing. Now spatial data visualization has many forms and these forms are more interactive and clear, some of them combining with data mining method are intelligent. Point data, line data and polygon data are three formats data that can be displayed in spatial visualization. About point data, user can use points locating on the map to indicate some place, but as the points begin to increase, put all the points on one map can make the map chaotic, at this time cluster the point or use density map is a good choice. Besides point data, line data is used to denote street and polygon data can be used to indicate an area. [Zhang and Li 2012] For example, Figure 27a shows the America Median income. In this map, dark color means high income and the light color means the low income.

TerraFly GeoCloud provides spatial analysis and the result visualization. Result visualization combines data visualization (histogram, pie chart and so on) and geospa-



(a) America Median Income (b) Customized map

Fig. 27: Related work

tial data visualization, and the visualization form is interactive and intelligent. The spatial data visualization gives the data more comprehensive view.[Rishe et al. 2004]

7.2. Spatial analysis

Spatial analysis is especially used on geographic data. The difference between spatial analysis and traditional analysis is that spatial analysis methods use spatial information of the data, such as the location, orientation and adjacent areas. Spatial analysis is widely used in many domains which contain biology, ecology, epidemiology, ecology and criminology, etc. There are some kinds of spatial analysis methods which include spatial clustering, spatial autocorrelation, spatial regression, spatial interpolation and spatial distribution measurement.[Fotheringham and Rogerson 2013]

As increasing data , developing GIS system and advancing analytic methods bring new opportunities for applying analytic methods and displaying the results of spatial analysis. TerraFly GeoCloud presents comprehensive spatial analysis methods and result visualization in a more interactive way. User can leverage these methods without programming, and get the result visualized on the map with a few clicks.[Bailey et al. 1994]

7.3. Customized map visualization

Nowadays more and more information contains geographic component, but how to visualize geographic component in a rapid and customized way receives less interest in science. But the demand and meaning of customized map visualization is tremendous. Most of the data includes geographic information , such as demographic data, socioeconomics data and biological data. These data can effect the new policy of government, new discovery of science and new strategy of company. For example, for a business company, sales(where sales is good and where sales is bad) data or target customer location needs to be shown to adjust company's plan.

Customized map visualization meet several challenges.First, hard to get rapid and accurate map.User needs to use complicated program to get map from traditional map visualization software. Second, hard to get really customized map. Some map service can provide some customized view for user. For example, figure 27b shows near adjacent data are merged together and use a big circle to indicate these data, or cluster the data. But it can not allow user to manipulate the data in their own way, because there are only several the map visualization styles provided.

TerraFly Geocloud provides MapQL as a spatial query and map render tool. User can query and visualize the data use a SQL-like statements. Because Geocloud is a web-based online service, user can use MapQL online and get a result in the map directly. This SQL-like statements facilitate users and let them draw the map in their own ways. [Teng et al. 2006][Wang 2011]

7.4. related products

In the geospatial discipline, web-based GIS services can significantly reduce the data volume and required computing resources at the end-user side.[Li et al. 2010][Fotheringham and Rogerson 2013] To the best of our knowledge, TerraFly GeoCloud is one of the first systems to study the integration of online visualization of spatial data, data analysis modules and visualization customization language.

Various GIS analysis tools are developed and visualization customization languages have been studied in the literature. ArcGIS is a complete, cloud-based, collaborative content management system for working with geographic information. But systems like ArcGIS and Geoda focus on the content management and share, not online analysis.[Johnston et al. 2001][Anselin et al. 2006] Azavea has many functions such as optimal Location find, Crime analysis, data aggregation and visualization. It is good at visualization, but has very limited analysis functions.[Boyer et al. 2011]

Various types of solutions have been studied in the literature to address the problem of visualization of spatial analysis. However, on one hand, good analysis visualization tools like Geoda and ArcGIS do not have online functions. To use them, users have to download and install the software tools, and download the datasets. On the other hand, good online GIS systems like Azavea, SKE, and GISCloud have limited analysis functions. Furthermore, none of above products provides a simple and convenient way like MapQL to let user create their own map visualization.[Hearnshaw et al. 1994][Boyer 2010] The related products are summarized in Table I. Our work is complementary to the existing works and our system also integrates the data mining and visualization.

Table I: GIS Visualization Products

Name	Website	Product features description	Product features description
ArcGIS Online	http://www.arcgis.com	http://www.arcgis.com ArcGIS Online is a complete, cloud-based, collaborative content management system for working with geographic information.	No online Analysis, focus on the content management and share.
Azavea	http://www.azavea.com/products/	optimal Location find, Crime analsis, data aggregated and visualized	Good visualization. Very limited Analysis functions
SKE	http://www.skeinc.com/GeoPortal.html	Spatial data Viewer	Focus on the spatial data viewer.
GISCloud	http://www.giscloud.com	with few analysis (Buffer , Range , Area , Comparison , Hotspot , Coverage , Spatial Selection)	Very limited simple analysis.
GeoIQ	http://geocommons.com/	filtering, buffers, spatial aggregation and predictive	Focus on GIS, very good Visualization and interactive operation. Very limited and simple analysis: currently provide predictive (Pearsons Correlation).

ACKNOWLEDGMENTS

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

REFERENCES

Luc Anselin. 1995. Local indicators of spatial associationLISA. *Geographical analysis* 27, 2 (1995), 93–115.

Luc Anselin, Ibnu Syabri, and Youngihn Kho. 2006. GeoDa: an introduction to spatial data analysis. *Geographical analysis* 38, 1 (2006), 5–22.

Peter Armitage, Geoffrey Berry, and John Nigel Scott Matthews. 2008. *Statistical methods in medical research*. John Wiley & Sons.

Trevor C Bailey, S Fotheringham, and P Rogerson. 1994. A review of statistical spatial analysis in geographical information systems. *Spatial analysis and GIS* (1994), 13–44.

Michel Bilodeau, Fernand Meyer, Michel Schmitt, and Georges Matheron. 2005. *Space, Structure and Randommess: Contributions in Honor of Georges Matheron in the Field of Geostatistics, Random Sets and Mathematical Morphology*. Springer.

Deborah Boyer. 2010. From internet to iPhone: providing mobile geographic access to Philadelphia's historic photographs and other special collections. *The Reference Librarian* 52, 1-2 (2010), 47–56.

Deborah Boyer, Robert Cheetham, and Mary L Johnson. 2011. Using GIS to Manage Philadelphia's Archival Photographs. *American Archivist* 74, 2 (2011), 652–663.

HJ De Knegt, F Van Langevelde, MB Coughenour, AK Skidmore, WF De Boer, IMA Heitkonig, NM Knox, R Slotow, C Van der Waal, and HHT Prins. 2010. Spatial autocorrelation and the scaling of species-environment relationships. *Ecology* 91, 8 (2010), 2455–2465.

Robin Dubin, R Kelley Pace, and Thomas G Thibodeau. 1999. Spatial autoregression techniques for real estate data. *Journal of Real Estate Literature* 7, 1 (1999), 79–96.

Paul Elliott and Daniel Wartenberg. 2004. Spatial epidemiology: current approaches and future challenges. *Environmental health perspectives* (2004), 998–1006.

Martin Ester, Hans-Peter Kriegel, J Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *KDD*, Vol. 96. 226–231.

Stewart Fotheringham and Peter Rogerson. 2013. *Spatial analysis and GIS*. CRC Press.

Arthur Getis and J Keith Ord. 1992. The analysis of spatial association by use of distance statistics. *Geographical analysis* 24, 3 (1992), 189–206.

Hilary M Hearnshaw, David John Unwin, and others. 1994. *Visualization in geographical information systems*. John Wiley & Sons Ltd.

Kevin Johnston, Jay M Ver Hoef, Konstantin Krivoruchko, and Neil Lucas. 2001. *Using ArcGIS geostatistical analyst*. Vol. 380. Esri Redlands.

Harry H Kelejian and Ingmar R Prucha. 1998. 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.

Martin Kulldorff. 1997. A spatial scan statistic. *Communications in Statistics-Theory and methods* 26, 6 (1997), 1481–1496.

Martin Kulldorff and Neville Nagarwalla. 1995. Spatial disease clusters: detection and inference. *Statistics in medicine* 14, 8 (1995), 799–810.

Alvin CK Lai, Tracy L Thatcher, and William W Nazaroff. 2000. Inhalation transfer factors for air pollution health risk assessment. *Journal of the Air & Waste Management Association* 50, 9 (2000), 1688–1699.

Hongfei Li, Catherine A Calder, and Noel Cressie. 2007. Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. *Geographical Analysis* 39, 4 (2007), 357–375.

Xiaoyan Li, Liping Di, Weiguo Han, Peisheng Zhao, and Upendra Dadi. 2010. Sharing geoscience algorithms in a Web service-oriented environment (GRASS GIS example). *Computers & Geosciences* 36, 8 (2010), 1060–1068.

Yun Lu, Mingjin Zhang, Tao Li, Yudong Guang, and Naphtali Rishe. 2013. Online spatial data analysis and visualization system. In *Proceedings of the ACM SIGKDD Workshop on Interactive Data Exploration and Analytics*. ACM, 71–78.

Nathan Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer research* 27, 2 Part 1 (1967), 209–220.

Patrick AP Moran. 1950. Notes on continuous stochastic phenomena. *Biometrika* 37, 1-2 (1950), 17–23.

L John Old. 2002. Information Cartography: Using GIS for visualizing non-spatial data. In *Proceedings, ESRI International Users' Conference, San Diego, CA*.

Stan Openshaw, Martin Charlton, Colin Wymer, and Alan Craft. 1987. 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.

Naphtali Rishe, Shu-Ching Chen, Nagarajan Prabakar, Mark Allen Weiss, Wei Sun, Andriy Selivonenko, and D Davis-Chu. 2001. TERRAFLY: A High-Performance Web-based Digital Library System for Spatial Data Access.. In *ICDE Demo Sessions*. 17–19.

N Rishe, M Gutierrez, A Selivonenko, and S Graham. 2005. TerraFly: A tool for visualizing and dispensing geospatial data. *Imaging Notes* 20, 2 (2005), 22–23.

Naphtali Rishe, Yanli Sun, Maxim Chekmasov, Andriy Selivonenko, and Scott Graham. 2004. System architecture for 3D terrafly online GIS. In *Multimedia Software Engineering, 2004. Proceedings. IEEE Sixth International Symposium on*. IEEE, 273–276.

Robert Spence and A Press. 2000. Information visualization. (2000).

Michael L Stein. 1999. *Interpolation of spatial data: some theory for kriging*. Springer.

William Teng, Naphtali Rishe, and Hualan Rui. 2006. Enhancing access and use of NASA satellite data via TerraFly. In *Proceedings of the ASPRS 2006 Annual Conference*.

Jon Wakefield and Paul Elliott. 1999. Issues in the statistical analysis of small area health data. *Statistics in medicine* 18, 17-18 (1999), 2377–2399.

Huan Wang. 2011. A large-scale dynamic vector and raster data visualization geographic information system based on parallel map tiling. (2011).

Yi Zhang and Tao Li. 2012. DClusterE: A Framework for Evaluating and Understanding Document Clustering Using Visualization. *ACM Transactions on Intelligent Systems and Technology (TIST)* 3, 2 (2012), 24.

Sagit Zolotov, Dafna Ben Yosef, Naphtali D Rishe, Yelena Yesha, and Eddy Karnieli. 2011. Metabolic profiling in personalized medicine: bridging the gap between knowledge and clinical practice in Type 2 diabetes. *Personalized Medicine* 8, 4 (2011), 445–456.

Received February 2007; revised March 2009; accepted June 2009

Online Appendix to: Online Spatial Data Analysis and Visualization System

Mingjin Zhang, Florida International University
Huibo Wang, Florida International University
Yun Lu, Florida International University
Tao Li, Florida International University
Yudong Guang, Florida International University
Chang Liu, Florida International University
Erik Edrosa, Florida International University
Naphtali Rishe, Florida International University

DOI: http://dx.doi.org/10.1145/0000000.0000000