Technical communication

Development of SOVAT: A numerical–spatial decision support system for community health assessment research

Matthew Scotch\textsuperscript{a}, Bambang Parmanto\textsuperscript{a,b,*}

\textsuperscript{a} Center for Biomedical Informatics, University of Pittsburgh, PA 15260, USA
\textsuperscript{b} Department of Health Information Management, University of Pittsburgh, Pittsburgh, PA 15260, USA

Abstract

Introduction: The development of numerical–spatial routines is frequently required to solve complex community health problems. Community health assessment (CHA) professionals who use information technology need a complete system that is capable of supporting the development of numerical–spatial routines.

Background: Currently, there is no decision support system (DSS) that is effectively able to accomplish this task as the majority of public health geospatial information systems (GIS) are based on traditional (relational) database architecture. On-Line Analytical Processing (OLAP) is a multidimensional data warehouse technique that is commonly used as a decision support system in standard industry. OLAP alone is not sufficient for solving numerical–spatial problems that frequently occur in CHA research. Coupling it with GIS technology offers the potential for a very powerful and useful system.

Methodology: A community health OLAP cube was created by integrating health and population data from various sources. OLAP and GIS technologies were then combined to develop the Spatial OLAP Visualization and Analysis Tool (SOVAT).

Results: The synergy of numerical and spatial environments within SOVAT is shown through an elaborate and easy-to-use drag and drop and direct manipulation graphical user interface (GUI). Community health problem-solving examples (routines) using SOVAT are shown through a series of screen shots.

Discussion: The impact of the difference between SOVAT and existing GIS public health applications can be seen by considering the numerical–spatial problem-solving examples. These examples are facilitated using OLAP–GIS functions. These functions can be mimicked in existing GIS public applications, but their performance and system response would be significantly worse since GIS is based on traditional (relational) backend.

Conclusion: OLAP–GIS system offer great potential for powerful numerical–spatial decision support in community health analysis. The functionality of an OLAP–GIS system has been shown through a series of example community health numerical–spatial problems. Efforts are now focused on determining its usability during human–computer interaction (HCI). Later work will focus on performing summative evaluations comparing SOVAT to existing decision support tools used during community health assessment research.

© 2005 Elsevier Ireland Ltd. All rights reserved.

* Corresponding author at: Department of Health Information Management, 6026 Forbes Tower, University of Pittsburgh, Pittsburgh, PA 15260, USA. Tel.: +1 412 383 6649; fax: +1 412 383 6655.
E-mail addresses: scotch@cbmi.pitt.edu (M. Scotch), parmanto@pitt.edu (B. Parmanto).
1386-5056/$ – see front matter © 2005 Elsevier Ireland Ltd. All rights reserved.
1. Introduction

The development of numerical–spatial routines is frequently required to solve complex community health problems. Community health assessment (CHA) professionals who use information technology need a complete system that is capable of supporting the development of numerical–spatial routines. Currently, there is no decision support system (DSS) that is effectively able to accomplish this task and thus decision makers are forced to improvise these steps with individual software applications.

In many problem-solving instances, the individual is a relatively experienced person who uses past experiences of solving similar problems and applies similar problem-solving techniques to the current dilemma. In general, a routine is the typical steps the individual uses for solving a given problem. There will be two types of steps highlighted: numerical and spatial. A numerical step only involves numerical data, while a spatial step involves spatially defined data (for example objects that have a coordinate value). For the purposes of this paper, a numerical–spatial routine is one that contains both numerical and spatial steps and attempts to solve numerical–spatial problems. As an example, a typical routine during community health assessment involves both spatial and numerical steps, and can be described as such:

1. identify geographic community of interest;
2. identify health factors within the community;
3. identify bordering communities of interest;
4. identify health factors within bordering communities;
5. compare factors within community against factors of bordering community;
6. identify aggregate (statewide, or national, etc.) community;
7. identify health factors within aggregate community;
8. compare factors within community against factors of aggregate community.

The first step, identification of a geographic community is spatial. For example, let’s say the community in question is a county. This can be satisfied by clicking on a county in a digital map using geospatial information systems (GIS). This step represents the act of merely signifying the area or region of interest. The second step, identifying the health factors within the community, is purely numerical. For example, the ranking of top five inpatient circulatory system diseases per 1000 for a particular age category aggregated at the community level is a numerical process. This can be accomplished by querying a database or a data warehouse such as OLAP. However, the next step, identifying the bordering communities of interest, is purely spatial. For example, this can be done in GIS by highlighting the bordering counties on a digital map. The identification of health factors in these counties is purely numerical as in step 2. This time, however, the researcher might decide to aggregate the individual counties into a single set (for example bordering county A + bordering county B + bordering county C = Set S) for comparison purposes. The comparison between the bordering communities (Set S) and the original community is completely numerical. This same process can be continued to the aggregate level. The selection of a higher level of geographic granularity is spatial, while the calculation of disease ranking serves is a numerical step.

We have developed the Spatial OLAP Visualization and Analysis Tool (SOVAT) to address this type of numerical–spatial problem solving for the community health assessment domain [1]. This novel system is comprised of two core technologies: On-Line Analytical Processing (OLAP) and geospatial information system (GIS). This paper will discuss the development of SOVAT and demonstrate its usefulness by solving a series of numerical–spatial community health assessment problems.

2. Background

Table 1 shows the technologies frequently used in the development of decision support systems for supporting spatial or numerical problem solving, as well as the important decision support functions they provide.

As evident, most of these address some, but not all of the essential components for numerical–spatial problem solving. There are a few reasons for why a comprehensive system does not yet exist. One is that each technology has their own focus, and thus other features were not deemed important. For example, the purpose of statistical software packages is to allow the user to perform statistical analysis, and thus little emphasis is placed on the management of large, complex data sets. Another reason for why a comprehensive numerical–spatial system does not exist lies in the difficulty of providing a system that interfaces together spatial and numerical functions. For example, most off-the-shelf data mining software performs numerical data mining functions and thus the focus is on the discovery of patterns among numerical data. A system like this would thus have difficulty with the coupling of spatial functions, such as geo-coding or color gradation of spatial objects within a coordinate system.

<table>
<thead>
<tr>
<th>Table 1 – DSS technologies used for decision support and the functions they provide</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Interesting patterns/knowledge discovery</strong></td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Statistical software</td>
</tr>
<tr>
<td>Data mining</td>
</tr>
<tr>
<td>GIS</td>
</tr>
<tr>
<td>Relational DB OLAP</td>
</tr>
<tr>
<td>SOVAT</td>
</tr>
</tbody>
</table>
GIS software supports spatial display and spatial analysis. In addition it might provide statistical functions either by using plug-ins or by linking it with statistical software. Likewise relational database systems can be linked to statistical packages. However, the functionalities are usually limited or they require extensive customizing. For example, ESRI’s ArcGIS Geostatistical Analyst is a statistical extension to ArcView and other ArcGIS products [2]. These extensions allow for kriging, spatial correlation and probability maps to be developed from relational ArcGIS data tables. However, they are limited in that they do not offer the core functions available in stand-alone statistical software. As an example, data mining technologies such as Clementine and IBM Intelligent Miner have sufficient statistical functions because they entail the core properties of a full-featured stand-alone software package. ArcView was not designed to support core functions of powerful statistical packages.

By combining OLAP, GIS, and data mining, Table 1 shows our SOVAT system contains all the necessary functionality for a powerful numerical–spatial DSS. Data mining can complement the synergy of and OLAP–GIS system even more by allowing for the discovery of interesting patterns through classification of spatial attributes, such as counties or municipalities. However, the two key technologies that together form the core of a powerful numerical–spatial DSS are OLAP and GIS. These two technologies will now be briefly discussed.

2.1. OLAP

On-Line Analytical Processing (OLAP) supports multidimensional data warehouse modeling. OLAP allows for rapid queries of multidimensional data that enables for powerful analysis and discovery through visual display on easy-to-use graphical user interfaces (GUI). With OLAP, data is represented conceptually as a multidimensional cube which enables the user to view different dimensions of multiple datasets and then query several dimensions at once.OLAP supports several distinct functions for data retrieval and analysis such as: drill-up (decreasing granularity, e.g., United States data by month → United States data by year), drill-down (increasing granularity, e.g., United States data by year → United States data by month), and slice and dice (retrieving a sub-section of data, e.g., Data for May and June for only New Hampshire and Vermont). All of these functions act on the multidimensional data cube and are performed almost instantaneously.

While OLAP is recognized as a useful component within today’s decision support framework for numerical problem solving, its potential has yet to be realized in the realm of spatial problem solving. Combining OLAP with other technologies to create a numerical–spatial framework could significantly enhance the usefulness of numerical–spatial problem solving and lead to better decision making. The reasons for why OLAP has not been integrated with such technologies as GIS for decision support lie in the difficulty and ambiguity in combining numerical and spatial frameworks [3].

2.2. GIS

GIS is not considered a decision support system based on the fact that it lacks the ability for supporting problem specific modeling [3]. Attempts have been made to enhance querying capabilities with stand-alone GIS packages by focusing on such spatial analysis functions as network analysis and buffering, but this is still considered insufficient from a decision support perspective. GIS provides many useful functions for community health assessment spatial problem solving such as buffering, network analysis, overlaying, and clipping. While GIS is useful as a stand-alone product for geospatial analysis, its inflexibility in relation to numerical–spatial problem solving has proven a challenge for DSS developers. The issue seems to lie in what other technologies should be incorporated with GIS in order to create an effective support system for creating spatial and numerical steps. None of the previous decision support systems have adequately addressed this problem. We believe the foundation for an effective numerical–spatial decision support system is to combine GIS with On-Line Analytical Processing. Some examples of numerical–spatial systems will now be discussed.

2.3. Numerical–spatial decision support systems

Bedard et al. [4] are one of the few researchers who have combined OLAP and GIS into a decision support system. The authors developed the ICEM-SE project to integrate geographic knowledge discovery for Canadian environmental health surveillance. Recently, a fully deployed version of ICEM-SE called JMAP Spatial OLAP was introduced as a web-enabled Java application [5]. The user interface of ICEM-SE was developed using off-the-shelf plug-in technology provided by ProClarity,¹ which is a desktop client that provides presentation layer support through the display of numerical graphs and charts. The ICEM-SE architecture appears to combine all the data, both spatial and non-spatial, in a temporary data warehouse. This data is then used to process the OLAP cube. For a detailed look at the ICEM-SE architecture, we refer the readers to [4]. The combination of GIS and OLAP creates the potential for an enhanced decision support environment. The authors claim that the users will be able to:

- Use OLAP functionality such as roll-up/drill-down in order to navigate through different levels of detail amongst both spatial and non-spatial data.
- Switch easily from different health themes (asthma, cancer, drinking water quality).
- Produce different statistical measures ([6], p. 81).

The ability for one system to support all these functions is made possible through the combination of OLAP and GIS. A system built in this framework provides the user with a powerful environment for numerical–spatial problem solving. The author describes a problem-solving example in relation to identifying causes of asthma hospitalization. The interaction with the interface involves selecting dimension attributes such as a disease dimension (for the selection of ‘asthma’) to a (health) facilities dimension, a numerical measures dimension, and finally a temporal dimension. The attribute selection process is facilitated by the multidimensional architecture of OLAP. The user can interactively drill-up and or drill-down on

¹ http://www.proclarity.com (11/18/04).
specific dimensions in order to switch from one level of aggregation to another. The result is then displayed on the map.

The epidemiologic query and mapping system also known as EpiQMS, is an example of a web-based system that enables users to perform statistical analysis on public health data using different data views such as: tables, charts, and maps [6]. The system was developed by the University of Washington in an effort to increase the availability of up-to-date on-line health-related data to three different types of users: the general public, public health professionals, and clinicians [6]. One of the main focuses of the project is to secure data confidentiality through the creation of cell size rules that prohibit certain types of users from accessing data in a specific level of geographic aggregation that is above a certain threshold count. For example, the general public may have access to death statistics down to the census tract level with no threshold on cell sizes, but with more sensitive information such as sexually transmitted diseases (STDs), only have access down to the zip code level with cells having a count greater than five [6]. Different types of users, such as public health researchers who need more detailed level of analysis, might be able to access death statistics down to the block group with no cell threshold. Another manner in which EpiQMS deals with data confidentiality is to enforce “concurrent dimension” rules [6], p. 32). This stipulates how many cross tabulations can occur in a query. For example, if the number of concurrent dimensions was set to three, the user would only be able to query on three dimensions at a time (for example, ‘sex’ by ‘age’ by ‘race’). This technique obviously reduces the potential for small cell sizes by limiting the number of possible dimension combinations. EpiQMS contains many statistical measurements that are pre-calculated offline in SAS including age-adjusted rates, confidence intervals, standard mortality/morbidity rates (SMR), nearest neighbor, and Poisson probability mapping. As mentioned, the system contains a variety of output reports including tables, charts, maps, and area profiles. These reports may contain hospitalization data or death information for various age groups by sex and by race. The maps are produced using scalar vector graphics (SVG) which is a format for presenting spatial data on the web. EpiQMS does not offer spatial analysis; only providing presentation of digital maps. Thus it is offers no support in regards to spatial problem solving. The system appears to have a traditional database format. In addition, the mapping component is mostly a reporting medium rather than an interface feature that supports query formulation through end-user interaction.

Hernandez et al. [7] have developed the commonGIS system for environmental decision support in Nicaragua. This system is a GIS application that provides spatial analysis, display, and interactivity through an elaborate graphical user interface. It has been coupled with the 'SPIN!' system [8] which contains data mining functionality. The commonGIS system makes 'SPIN!' a powerful spatial data mining decision support system that can present both spatial analysis and numerical mining results simultaneously through the interface. Currently the system does not utilize OLAP; however, there are plans to add it to commonGIS in the near future [7].

The Nehme and Simoes project for land evaluation [9] is another system that uses the GIS component to simply display information rather than incorporating it into the process of query development. It examines soil fertility, water excess/deficiency, erosion susceptibility, and the required degree of soil tillage. The system incorporates artificial intelligence (AI) techniques through the use of decision rules and learning algorithms (Bayesian Networks) to automatically produce land evaluation. While the combination of AI with decision support has produced success in the area of automatic model selection [10–12], it offers limited aid in the course of query development (other than feedback capabilities). A similar project is The National Agricultural Decision Support System (NADSS) [13]. It is being developed at the University of Nebraska to support the decision making process for drought analysis. The system calculates daily drought indices from a combination of disparate data sources in an effort to perform exposure analysis.

Certain agencies of the US Government have also created spatially related decision support systems. The United States Health Resources and Services Administration (HRSA) has developed a geospatial data warehouse for use by the general public [14]. Anyone who clicks onto the site and uses the map tool can develop their own map with HRSA program locations, facility locations, and demographic/population data. The process for map development is similar to the ICEM-SE interface. Layers and objects can be added as well as numerical measures and demographic attributes. For example, through a series of attribute selections, the user can view a map of Males 15-17 by county in Pennsylvania (with hospitals represented as points). The system does not appear to incorporate OLAP on the backend. In addition, there is no spatial query development feature with the map nor does it support geospatial analysis in any fashion. It appears as though the only GIS functionality supported is for the systematic view of spatial and numerical data. In addition, the numerical analysis is also limited; containing only simple percents and counts. The Centers for Disease Control (CDC) and Prevention also has a similar product called The Interactive Atlas of Reproductive Health that uses ESRI (ArcIMS)2 software to allow users to log on and create maps concerning reproductive health issues, such as infant mortality, birth weight, and fertility [15]. It is also weak in query development features.

The National Cancer Institute (NCI) is developing a GIS database tool for their Long Island Breast Cancer Study Project [16,17]. Using the system, the researchers hope to increase their awareness of environmental causes of breast cancer. The system was developed using ESRI’s GIS ArcView software,3 which is part of their desktop GIS suite. The system includes data mining techniques such as numerical clustering (by incorporating SaTScan software, which was developed by Kulldorff et al. [18]), in an effort to find interesting patterns within the data.

Other than the work by Bedard et al., these systems attempted to solve numerical–spatial problems by incorporating GIS without OLAP. GIS is effective for spatial problem solving, but a framework for its integration as a core technol-

---

ogy component within a DSS has yet to be established. Coupling it with a data subsystem based on a traditional database architecture is not sufficient. For this reason, we decided to integrate GIS with the multidimensional environment of OLAP. The next section will detail the development of our SOVAT decision support system for community health assessment research.

3. Methodology

Data necessary for performing community health assessments (health and population data) was collected from three main sources: The Pennsylvania Department of Health, the Pennsylvania Healthcare Cost Containment Council, and The United States Census Bureau. The Department of Health provided cancer registry data, birth, and death statistics. The Healthcare Cost Containment Council provided inpatient and outpatient healthcare utilization data. The Census Bureau provided population and socioeconomic data. Dimensions were identified for the development of the star schema and include: Age, Birth weight, Diagnosis, Education, Race, Region, Sex, and Year.

The diagnosis dimension is based off of the World Health Organization’s International Classification of Disease 9-CM coding hierarchy. This contains numerous disease categories from broad classifications down to the individual disease code. The geography dimension represents data for the state of Pennsylvania. The spatial data was provided by the US Census Bureau. Data can be downloaded (in shape file format) from their website. The spatial data consisted of three tables, each representing different levels of geographic granularity: statewide, county level, and municipality level. The hierarchy goes from statewide data to individual counties, and then to their municipalities. The lowest (leaf) level data comprises specific county subdivision codes (a municipality might have more than one). The geography dimension is the most important dimension for a community health assessment decision support system. It dictates the level of granularity with which the expert can browse the data. We designed ours at the municipality level, since it represents the highest level of geographic granularity that all our data sets have in common. While Pennsylvania was used to model our geography dimension, it should be noted that the flexibility of our system allows for any type of geography to be used. For example, given nationwide data, the dimension can be easily changed to model: Country → States → Counties → Municipalities. As explained before, it is only the data subsystem component that needs to be altered. The final dimension that needs to be addressed is the Race dimension. The census bureau altered their definition of a ‘race’ in their 2000 report to include the option of selecting more than one racial category. In order to reflect this change, we created the Race dimension as slowly changing which enabled the data subsystem to model this real-world change. The concept of slowly changing dimensions will not be discussed and is beyond the scope of this paper. For further information, we refer the reader to [19].

Once the dimensions were defined, a star schema was developed to represent each of the seven data sets (cancer incidence, birth, death, inpatient hospitalization, outpatient hospitalization, population, and socioeconomic). Individual data cubelets were then created and then combined to form the ‘virtual’ community health OLAP cube. The tables were housed in the data subsystem component of SOVAT.

The combination of OLAP and GIS occurred through the development of an integration engine (created using the NET platform). The integration engine enables both the OLAP and spatial functions to act on both the numerical and the spatial data. The difficulty with combining numerical and spatial components is that the underlying data is vastly different, and the tools used to access these different types of data are vastly different. The software used for development is able to support many third party component software that utilizes both numerical and spatial functions, thus it was the logical choice as the development code for our integration engine.

The numerical–spatial interface was developed in NET platform. Two important DSS tools were added to allow for numerical–spatial problem solving. The spatial software consists of an entire suite of spatial analysis functions such as buffering and shortest path as well as the ability to display digital maps. The numerical software, which is used to display the results from the OLAP cube, allows for different types of graphs to be shown and changed on-the-fly in an interactive manner.

4. Results

We have developed the Spatial OLAP Visualization and Analysis Tool (SOVAT) system for numerical–spatial decision support in community health assessment research. The system combines OLAP and GIS technology to create a unique and powerful tool for answering numerical–spatial problems in community health assessment research. The architecture of the decision support system can be seen in Fig. 1.

At the implementation level, the spatial and numerical data can be combined in the same database application (SQL Server). At the conceptual level, the spatial and the numerical data are much different. In our architecture, we chose to represent this conceptual difference by separating the two data sources at the beginning and then combining them in the integration engine.

The model subsystem of SOVAT contains three separate features: spatial analysis, data mining, and construction of communities. Spatial analysis such as buffering could be a valuable feature in a community health decision support system. For example, a researcher could use buffering to identify the rural areas that fall within a specified distance of a hospital.

Fig. 1 – Architecture of our SOVAT system. This data-oriented DSS combines OLAP and GIS through the integration engine that enables users to construct numerical–spatial routines.

tal and examine specific disease incidences that dramatically increase outside these areas. Buffering is an available feature in SOVAT.

The ability to discover interesting patterns among mountains of health and population data is a valuable feature to a community health assessment expert. Data mining functions such as clustering would make this possible. For example, numerical clustering of geographical items (such as counties) would reveal which are similar based on a series of non-geographical attributes. This goes beyond using a health-related measure to show color gradation on a map.

The final component to the model subsystem is the ability to create a customized community. Many times, a community is more than one county or one municipality. A researcher could need to examine a series of counties or municipalities. Thus numerical aggregation needs to be done on these attributes to create a defined set or community that enables the researcher to examine these individual areas as one unit. For example, if a CHA researcher needed to examine Northwest Pennsylvania which was defined as Erie, Crawford and Warren counties, he/she would need a way to aggregate these areas together so they could classified as one region rather than three counties. SOVAT allows for communities to be easily defined.

4.1. SOVAT interface

To the user, the interface is the system. Thus, in order to realize the integration of OLAP and GIS technologies for numerical–spatial problem solving, it was important to develop an easy-to-use yet powerful interface. Fig. 2 shows the SOVAT GUI which supports drag and drop as well as direct manipulation actions.

The top left corner of the interface contains the dimensions that are stored in the system. The types of dimensions represent the combination of health and population data and are: Age, Birth weight, Diagnosis, Geography, Measures, Race, Region, Sex, and Year. Each dimension has its own tab. Dimension attributes are represented in a hierarchical (OLAP) fashion on each tab and can be explored by clicking on the ‘+’ or ‘−’ next to the attribute name. Next to the dimension tabs is the charting area. This displays the numerical results in either a bar chart or line graph format. This type of display is common in many OLAP front-end client applications. Below the chart is the map. The map displays the same numerical results as the chart above, but in the form of a color-gradated map. The map’s legend details the meaning of the color grades for each query. To the left of the charting area are the row, column, and background list boxes. The row and column list boxes define the nature of the query and require dimensions from the background list. This is another feature that is common in many OLAP front-end clients.

There are many different ways to submit a query using our SOVAT interface. The standard OLAP-like method is to select attributes from the dimension tabs and drag and drop them onto the charting area. Multiple attributes may be selected from the same or different dimension tabs by using the control or shift key. If the user gets tired of dragging and dropping, a query can also be performed by selecting the attributes and then selecting the ‘Query’ button by background list. Querying can also be done through direct manipulation of the chart or the map. Regardless of which is manipulated, the query results will be displayed in both presentations. To query with the map, the user may select a map element (such as a bar) and right click. This displays the OLAP functions, drill-up and down. Selection of either one will perform an action on that
specific cell. For example, if the user selects an Allegheny County bar, performing a drill-down will display the municipalities (or children) of that attribute.

Direct manipulation with the map is a little different. Multiple map items, such as counties, may be selected by holding down the shift or control buttons as they are clicked with the mouse. Double clicking on these elements will automatically perform a drill-down function. Thus to perform the same query as with the bar chart, Allegheny County can be selected on the map and then double clicked. A municipality-level map will then display its counties. Once highlighted, a right click will show the user additional options such as: ‘Drill-Up’, ‘Drill-Out’, ‘Buffering’, and ‘Save as Community.’ Choosing ‘Drill-Up’ will display a map with a lower level of geographic granularity. For example, a drill-up on a county will show a map of the entire state without county borders. Drill-out is a function that is not included in standard OLAP technology. Instead of increasing or decreasing the level of geographic granularity, drill-out “expands outward” and performs numerical aggregation on bordering areas. For example, a drill-up on a county will show a map of the entire state without county borders. Drill-out on the other hand combines GIS and OLAP features by first detecting the adjacent spatial objects that surround selected map items, and then using numerical aggregation on these identified bordering areas to allow for geographic comparison.

The final feature for map manipulation is the ‘Save as Community’ option. This modeling option enables the researcher to customize his/her own-defined community. As mentioned, a community may be more than one county or one municipality. Enabling the researcher to use the spatial information to customize a series of geographical areas into one set has the potential to facilitate problem solving. This process is done very easily with SOVAT. The user only needs to select the areas he/she wants to include in the community. Right clicking and selecting ‘Save as Community’ enables the user to create a name for their customized community. That community will then be added to the Community dimension tab. A query using this community would display results as if the entire area was one county or one municipality.

The final way to produce a query with SOVAT is the use of the custom function buttons to the right of the background dimension. The most prominent is the Top 5 function. This performs ranking analysis on the data and displays the top five results. This allows the community health researcher to quickly identify the leading health issues within a community. For example, using this function could display the top five causes of death or the leading causes of cancer hospitalization. Top 5 can also be performed from with a geographical

---

9 The Community dimension is a client-side dimension for modeling purposes and is not contained in the data subsystem like the other dimensions previously mentioned.
focus for example displaying the Top 5 counties that have outpatient cataract cases.

4.2. Problem solving with SOVAT

The following numerical–spatial problem-solving examples will highlight how SOVAT can be used for development of numerical–spatial routines in community health assessment research. A step-by-step approach will demonstrate the system functionality.

Problem #1: What is the cancer incidence/100,000 of colon cancer in Southwest PA for 2000 (Washington, Greene, Allegheny County for demonstration purposes)?

The illustrations for this problem are shown below (Figs. 3–6):
- Select the appropriate counties by clicking them on the map while holding down the shift or control button.
- Right click and select ‘Save as Community.’
- Enter the name for the community (e.g. SW PA) and hit ‘ok.’ This allows the user to add this community to the system. Fig. 4 shows this result as SW PA is added to the Community dimension tab.
- Select the other attributes that need to be included in the query by going to the dimension tab area of the interface, scrolling to the appropriate tab, and selecting the necessary attribute. For this query, in addition to the Community dimension, the Year, Diagnosis, and Measures tabs need to be accessed and the appropriate attribute selected from each (Fig. 5).
- Drag and drop the attributes onto to the charting area. The correct solution is then shown on the interface (Fig. 6).

Problem #2: How does the death rate in 1999 per 100,000 of ischemic heart diseases in Allegheny county compare to its bordering counties?

The illustrations for this problem are shown below (Figs. 7–9):
- Select the query attributes. For this example, they include the Year, Diagnosis, and Measures dimension. As with the first example, the attributes can be selected by accessing the appropriate dimension tab and selecting each node with the mouse (Fig. 7).
- Drag and drop the attributes onto to the charting area.
- Use direct map manipulation to select the appropriate county (Allegheny). This can be done by clicking ‘Allegheny County’ once on the map.
- Right click and select the ‘Drill-Out’ option (Fig. 8). The bordering counties will be automatically identified and shown in the results (Fig. 9).

Problem #3: What are the Top 5 municipalities in Erie County of inpatient/100,000 from digestive system diseases in 1998?
Fig. 4 – Adding the community to the system. Naming the community and selecting ‘ok’ will then add it to the ‘Community’ dimension tab (upper left).

Fig. 5 – Selection of appropriate attributes. For this task, attributes in the Year, Diagnosis and Measures dimensions need to be included in the query along with the community. Here, the specific disease ‘malignant neoplasm of colon’ is selected by iterating through the Diagnosis hierarchy.
Fig. 6 – Task completion. Once all the attributes are selected, dragging and dropping them onto the charting area will produce the result on both the bar chart and the map for this numerical–spatial task.

Fig. 7 – Selection of appropriate attributes. For this task, attributes in the Year, Diagnosis and Measures dimensions need to be included in the query along with Geography. Here, the specific year, 1999, is selected from the Year hierarchy.
Fig. 8 – Drill-out function. Once the attributes are included in the query, the drill-out function is used to identify the bordering areas and perform the numerical analysis. Here, Allegheny County is selected on the map. Right clicking and selecting 'Drill-Out' will perform this function.

Fig. 9 – Task completion. Drill-out is the ideal function to use for this numerical–spatial task. The bordering geographical areas are identified and included in the numerical analysis.
Fig. 10 – Selection of appropriate attributes. For this task, attributes in the Year, Diagnosis and Measures dimensions need to be included in the query along with Geography. Here, the specific measure 'Inpatient/1000' is selected from the Measures hierarchy.

Fig. 11 – Top 5 function. SOVAT's custom Top 5 displays the top five attributes given a numerical measure. With the Geography dimension in the column list, highlighting a geography attribute, in this case Erie County, will display the top five municipalities within this county that have the highest rate of inpatient/1000 for digestive system diseases in 1999.
The illustrations for this problem are shown below (Figs. 10 and 11).

- Select the appropriate attributes. For this example, they include the Year, Diagnosis, and Measures dimension (Fig. 10).
- Drag and drop the attributes onto the charting area.
- Arrange the appropriate dimensions in the row and column list. The column list defines how the Top 5 function is carried out. For this example, Geography dimension needs to be the only dimension included in the column list, and thus it is necessary to remove the Year dimension and drop it back into the background list.
- Highlight the attribute in which to perform the Top 5 function, in this case, Erie County is the appropriate choice.
- Select the Top 5 button (Fig. 11).

5. Discussion

SOVAT provides the potential to be a very effective and powerful tool for numerical–spatial problem solving in community health assessment research. OLAP has been utilized as an effective tool in the business world for addressing numerical decision making. For community health assessment research, OLAP needs to be extended to incorporate spatial features and spatial analysis. GIS alone is not sufficient for this problem, since it lacks the capacity for addressing complex numerical problem solving. GIS would prove ineffective for addressing complex numerical–spatial problems such as the examples described in this paper. Besides the work of Bedard et al. [4], OLAP and GIS have yet to be combined for numerical–spatial problem solving. One of the main challenges in building the interface is to combine the multidimensional framework of OLAP with a transactional nature of GIS. Without the proper interface, creating an OLAP–GIS system will be futile. Users must realize the synergy of these two technologies through an easy-to-use GUI that presents the numerical and spatial information simultaneously and enables for direct manipulation of the data through both OLAP and geospatial analysis.

The existing applications described previously represent vastly different technologies to SOVAT. From an architecture perspective, other than Bedard’s system, they do not use OLAP on the backend (data subsystem). The impact of this difference can be seen by considering the problem-solving examples in the previous section. In these routines, problem solving is facilitated by using OLAP–GIS functions. These functions can be mimicked in existing GIS public applications, but their performance and system response would be significantly worse since GIS is based on traditional (relational) backend. For example, problem #2 requires “drilling-out” which is an OLAP–GIS function unique to SOVAT. For example, drill-out requires both spatial and numerical processes. First, spatial detection is required to identify bordering areas. Then numerical aggregation on these bordering areas is needed to perform geographic comparison. With large spatial data files and complex joins between tables, the time to completion of this problem-solving task could be greatly increased. The same problem can be seen by performing drill-down functions as is in the case in problem #3. This requires changing spatial granularity and performing retrieval of numerical results on this new and more detailed geographical level. This type of ad-hoc analysis could require scanning of large spatial data files and then complex joins between these files; all of which could significantly slow information retrieval. In SOVAT, the OLAP–GIS functions take seconds to perform since the data is fully materialized in the OLAP cube and the spatial functions allow for direct user manipulation during ad-hoc analysis.

6. Conclusion

Community health assessment researchers are forced to solve complex numerical–spatial problems with independent numerical or spatial technology. To address this problem, we have created the Spatial OLAP Visualization and Analysis Tool (SOVAT). This decision support system, combines Online Analytical Processing with geospatial information system (GIS) technology, and thus provides the potential for community health researchers to develop powerful and effective numerical–spatial routines. We have described the construction of SOVAT and detailed the components of its easy-to-use elaborate graphical user interface. We then demonstrated its functionality through a series of example community health numerical–spatial problems. We are in the process of performing think-aloud assessments in order to determine its usability during human–computer interaction (HCI). Once we have made the necessary improvements to deem the system ‘usable’ we will perform a summative evaluation comparing SOVAT to existing decision support tools used for community health assessment research. Comparing SOVAT to existing technologies such as popular off-the-shelf GIS and statistical software applications, as well as certain database technology, will hopefully demonstrate the potential of an OLAP–GIS system for community health problem solving.

Acknowledgements

The project is supported in part by The National Library of Medicine (NLM) Training Grant 5 T15 LM007059 to Matthew Scotch, and by the National Institute on Disability and Rehabilitation Research (NIDRR) and National Telecommunication and Information Administration (NTIA) to Bambang Parmanto. The authors would like to thank Dr. Ravi Sharm for input and feedback during the development of the system as well as I. Wayan Sugiantara for his programming work for SOVAT development. In addition, the authors would like to thank Mike Meit and the University of Pittsburgh’s Center for Rural Health Practice for funding support during the development process.

References


