

**AN OLAP-GIS SYSTEM FOR NUMERICAL-SPATIAL PROBLEM SOLVING IN  
COMMUNITY HEALTH ASSESSMENT ANALYSIS**

by

Matthew Laurence Scotch

BA, University of Rochester, 1998

MA, Columbia University, 2002

Submitted to the Graduate Faculty of

Medicine in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2006

UNIVERSITY OF PITTSBURGH

FACULTY OF MEDICINE

This dissertation was presented

by

Matthew Laurence Scotch

It was defended on

April 5<sup>th</sup> 2006

and approved by

Cindy Gadd, PhD, MBA

Valerie Monaco, PhD, MHCI

Ravi K. Sharma, PhD

Valerie Watzlaf, PhD

Bambang Parmanto, PhD  
Dissertation Director

Copyright © by Matthew Scotch  
2006

**AN OLAP-GIS SYSTEM FOR NUMERICAL-SPATIAL PROBLEM SOLVING IN  
COMMUNITY HEALTH ASSESSMENT ANALYSIS**

Matthew Laurence Scotch, PhD

University of Pittsburgh, 2006

Community health assessment (CHA) professionals who use information technology need a complete system that is capable of supporting numerical-spatial problem solving. On-Line Analytical Processing (OLAP) is a multidimensional data warehouse technique that is commonly used as a decision support system in standard industry. Coupling OLAP with Geospatial Information System (GIS) offers the potential for a very powerful system. For this work, OLAP and GIS were combined to develop the Spatial OLAP Visualization and Analysis Tool (SOVAT) for numerical-spatial problem solving.

In addition to the development of this system, this dissertation describes three studies in relation to this work: a usability study, a CHA survey, and a summative evaluation.

The purpose of the usability study was to identify human-computer interaction issues. Fifteen participants took part in the study. Three participants per round used the system to complete typical numerical-spatial tasks. Objective and subjective results were analyzed after each round and system modifications were implemented. The result of this study was a novel OLAP-GIS system streamlined for the purposes of numerical-spatial problem solving.

The online CHA survey aimed to identify the information technology currently used for numerical-spatial problem solving. The survey was sent to CHA professionals and allowed for them to record the individual technologies they used during specific steps of a numerical-spatial

routine. In total, 27 participants completed the survey. Results favored SPSS for numerical-related steps and GIS for spatial-related steps.

Next, a summative within-subjects crossover design compared SOVAT to the combined use of SPSS and GIS (termed SPSS-GIS) for numerical-spatial problem solving. Twelve individuals from the health sciences at the University of Pittsburgh participated. Half were randomly selected to use SOVAT first, while the other half used SPSS-GIS first. In the second session, they used the alternate application. Objective and subjective results favored SOVAT over SPSS-GIS. Inferential statistics were analyzed using linear mixed model analysis. At the .01 level, SOVAT was statistically significant from SPSS-GIS for satisfaction and time ( $p < .002$ ).

The results demonstrate the potential for OLAP-GIS in CHA analysis. Future work will explore the impact of an OLAP-GIS system in other areas of public health.

## **ACKNOWLEDGEMENTS**

I would like to thank my dissertation advisor Dr. Bambang Parmanto for his guidance and belief in my abilities as a doctoral student. For four great years his tutelage was instrumental in my growth as a researcher. I am proud to say he was not only an excellent advisor, but an excellent friend as well. Being paired with Dr. Parmanto, I always felt I was the luckiest doctoral student in the Center for Biomedical Informatics.

Thanks also to Dr. Valerie Monaco for her guidance during the usability study and the evaluation study. I really appreciate all of the time she put into our Shadyside brainstorming sessions. She quickly became one of my favorite Biomedical Informatics faculty members. It was a great experience working with her.

I would also like to thank the other members of my dissertation committee, including Dr. Cindy Gadd, Dr. Ravi Sharma, and Dr. Valerie Watzlaf for their valuable contribution to my dissertation. I consider myself very fortunate to have such a great committee. Dr. Gadd, thank you also for being such a great academic advisor during your tenure in CBMI. I always felt very lucky to have you.

Special thanks to Wayan Sugiantara for his hard work in fixing all of the usability bugs left behind from my code. He was patient and very reliable. I always knew I could count on him to fix the problem at hand. Thanks also to Stephanie Hackett for her help on the survey and during my evaluation study. It was very much appreciated.

Thank you also to my funding source: the National Library of Medicine. It is truly an extremely generous award they provide. Without their financial assistance, I would not have the opportunity to reach my goal of becoming a PhD.

Finally, I can't express in words how much the love of my family means to me. It is hands down the most important thing in my life. I am so thankful to have such an amazing family. We are such a tight unit and for that I am so lucky. Mom, dad, Adam, and Molly, you are my inspiration. Thank you for believing in me.

## TABLE OF CONTENTS

1.	INTRODUCTION .....	1
1.1.	Introduction.....	1
1.2.	Decision Support Systems (DSS) .....	4
1.3.	Problem Solving Routines by Type of DSS .....	7
1.4.	Issues with Data-oriented Decision Support for Development of Routines .....	8
1.5.	On-Line Analytic Processing.....	10
1.6.	Geospatial Information Systems .....	12
1.7.	Research Focus .....	13
2.	BACKGROUND .....	16
2.1.	Overview .....	16
2.2.	Technologies for Supporting Spatial and Numerical Problem Solving in DSS .....	18
2.2.1.	Statistical Software .....	19
2.2.2.	OLAP .....	21
2.2.3.	GIS .....	23
2.2.4.	Data Mining/Knowledge Discovery .....	25
2.3.	Decision Support Systems for Numerical-Spatial Problem Solving .....	28
3.	COMBINING OLAP AND GIS FOR DECISION SUPPORT .....	38
3.1.	Introduction and Background .....	38
3.2.	Development of SOVAT .....	44
3.2.1.	The Integration Engine .....	46
3.2.2.	OLAP and GIS Interface.....	46
3.3.	Features in SOVAT for Numerical-Spatial Problem Solving.....	47
3.3.1.	Knowledge Discovery.....	48
3.3.2.	The Handling of Large, Complex Data Sets.....	48
3.3.3.	Multidimensional Data Exploration.....	49
3.3.4.	Statistical Analysis.....	49
3.3.5.	Spatial and Numerical Presentation.....	50
3.3.6.	Spatial Analysis .....	51
3.3.7.	Numerical-Spatial Routines.....	52
4.	USABILITY ASSESSMENT OF SOVAT .....	53
4.1.	Introduction.....	53
4.2.	Background.....	53
4.3.	SOVAT Interface .....	54
4.4.	Methodology .....	57
4.4.1.	Recruitment and Setting.....	57
4.4.2.	Study Procedures .....	58
4.4.3.	Objective Measurements.....	59
4.4.4.	Subjective Measurements .....	60
4.5.	Results.....	60
4.5.1.	Interface Changes Made .....	61

4.5.1.1.	Dimension Tabs .....	61
4.5.1.2.	Chart Display .....	66
4.5.1.3.	Map Display .....	68
4.5.1.4.	Row/Column, and Special Functions.....	75
4.5.2.	Best and Worst Aspects of the Interface.....	82
4.5.3.	Subjective and Objective Results.....	84
4.6.	Discussion .....	88
4.7.	Limitations .....	92
4.8.	Conclusion .....	92
5.	INFORMATION TECHNOLOGY UTILIZATION IN COMMUNITY HEALTH ASSESSMENT ANALYSIS .....	94
5.1.	Introduction.....	94
5.1.1.	Community Health Assessment.....	94
5.1.2.	Initiators of Community Health Assessments.....	95
5.1.3.	Components within a Community Health Assessment.....	97
5.2.	Background .....	98
5.3.	Methodology .....	102
5.3.1.	Data Collection .....	104
5.4.	Results.....	105
5.5.	Discussion .....	111
5.5.1.	Public Health Curriculum .....	113
5.6.	Limitations .....	114
5.7.	Conclusion .....	114
6.	EVALUATION OF SOVAT FOR NUMERICAL-SPATIAL PROBLEM SOLVING IN COMMUNITY HEALTH ASSESSMENT RESEARCH .....	117
6.1.	Introduction.....	117
6.2.	Background.....	118
6.3.	Methodology .....	120
6.3.1.	Pilot Study.....	122
6.3.2.	Recruitment and Setting.....	123
6.3.3.	Software used in the Study .....	123
6.3.4.	Study Procedures .....	124
6.3.5.	Objective Measurements.....	125
6.3.6.	Subjective Measurements .....	126
6.3.7.	Statistical Analysis.....	126
6.4.	Results.....	127
6.4.1.	Objective Measurements.....	127
6.4.1.1.	Time .....	127
6.4.1.2.	Success Rate.....	129
6.4.2.	Subjective Measurements .....	131
6.4.2.1.	User Preference .....	133
6.4.3.	Mixed Model Analysis.....	136
6.4.3.1.	Time .....	136
6.4.3.2.	User Satisfaction .....	137
6.5.	Discussion .....	139
6.6.	Limitations .....	140

6.7.	Conclusion .....	141
7.	SUMMARY AND FUTURE DIRECTIONS .....	144
7.1.	Summary .....	144
7.2.	Future Directions .....	145
7.2.1.	OLAP-GIS for other domains .....	145
7.2.2.	Cognitive Issues in relation to an OLAP-GIS system .....	146
APPENDIX A:	Recruitment Flier for Usability Study .....	147
APPENDIX B:	Recruitment Letter for Usability Study .....	148
APPENDIX C:	Informed Consent Form for Usability Study .....	149
APPENDIX D:	Computer Background Questionnaire .....	154
APPENDIX E:	Tasks for Usability Study .....	156
APPENDIX F:	PSSUQ Questionnaire .....	161
APPENDIX G:	Open-ended Satisfaction Questionnaire for Usability Study .....	165
APPENDIX H:	Best and Worst Results from Usability Study .....	166
APPENDIX I:	Objective and Subjective Results from Usability Study .....	167
APPENDIX J:	CHA Survey .....	168
APPENDIX K:	Recruitment Email for CHA Survey .....	176
APPENDIX L:	Results from CHA Survey .....	177
APPENDIX M:	Apriori Algorithm for CHA Survey Results .....	179
APPENDIX N:	Informed Consent Form for Evaluation Study .....	180
APPENDIX O:	Tasks for Evaluation Study .....	185
APPENDIX P:	Recruitment Flier for Evaluation Study .....	195
BIBLIOGRAPHY	.....	196

## LIST OF TABLES

Table 2-1: DSS generators for decision support and the capabilities they provide. ....	19
Table 2-2: Data mining functions for OLAM. ....	26
Table 2-3: Comparison of the OLAP-GIS DSSs. ....	37
Table 3-1: Traditional DSS generators versus an OLAP-GIS system. ....	42
Table 4-1: Worst aspects of the SOVAT system from round 1 and round 5. ....	83
Table 4-2: Best aspects of the SOVAT system from round 1 and round 5. ....	84
Table 4-3: Time (in min) summarized by round. ....	88
Table 4-4: Answers summarized by round. ....	88
Table 4-5: The number of wins when comparing each round. ....	88
Table 5-1: Twenty seven survey participants by type of organization. ....	106
Table 5-2: Problem solving category and the most popular IT for that category. ....	108
Table 5-3: Technology for numerical problem solving steps. ....	109
Table 5-4: Technology for spatial problem solving steps. ....	109
Table 5-5: IT applications used with statistical software for community health assessment. ...	110
Table 5-6: IT applications used with GIS software use for community health assessment. ....	110
Table 6-1: Mean time (rounded to the nearest minute) and 99% CI per period per task. ....	127
Table 6-2: Mean time (rounded to the nearest minute) and 99% CI per task. ....	129
Table 6-3: Success rate for the tasks. ....	130
Table 6-4: Mean Satisfaction scores and 99% CI per period. ....	131
Table 6-5: Mean Satisfaction scores and 99% CI. ....	132
Table 6-6: Positive responses in relation to SOVAT during the post-study interview. ....	134
Table 6-7: Negative responses in relation to SOVAT during the post-study interview. ....	136
Table 6-8: Mixed model analysis of Time variable. Shown are p-values per effect per task...	136
Table 6-9: Mixed model analysis of User Satisfaction. ....	137

## LIST OF FIGURES

Figure 1-1: Results showing descriptive counts of cholera deaths. ....	3
Figure 1-2: Numerical-spatial problem solving.....	4
Figure 1-3: An example of a traditional DSS architecture. ....	5
Figure 1-4: A Multidimensional OLAP Cube. ....	11
Figure 1-5: An example of GIS application (here using ArcGIS). ....	13
Figure 1-6: SOVAT architecture. ....	14
Figure 3-1: "Drill-Out" function in an OLAP-GIS system. ....	39
Figure 3-2: "Drill-Out" for boundary detection.....	39
Figure 4-1: Original SOVAT interface at 'startup'.....	54
Figure 4-2: New dimension tabs for round 2 with added search option. ....	62
Figure 4-3: Dimension tabs for round 4 of the usability study. ....	65
Figure 4-4: SOVAT chart for round 1 of the usability study.....	67
Figure 4-5: SOVAT chart for round 2 of the usability study.....	68
Figure 4-6: SOVAT map for round 2 of the usability study.....	69
Figure 4-7: SOVAT map for round 3 of the usability study.....	70
Figure 4-8: Help message when a community-related query is submitted. ....	71
Figure 4-9: SOVAT map used in round 4 of the usability analysis.....	72
Figure 4-10: SOVAT map for round 5 of the usability study.....	73
Figure 4-11: Row/Column error from round 1 of the usability study.....	76
Figure 4-12: Top 5 Wizard implemented in round 2 of the usability study. ....	78
Figure 4-13: Row/Column and special features component for round 2 of the usability study..	79
Figure 4-14: Special Function components for round 3 of the usability study.....	79
Figure 4-15: Special Function components for round 4 of the usability study.....	81
Figure 4-16: Final version of SOVAT interface after round 5 of the usability study.....	82
Figure 4-17: PSSUQ results summarized by round.....	85
Figure 4-18: Erroneous actions summarized by round. ....	86
Figure 4-19: Problem space occurrences summarized by round. ....	87
Figure 5-1: Example responses from Part 2 of the survey.....	107
Figure 5-2: CHA professional who uses both GIS and statistical software.....	112
Figure 5-3: Same participant and the response for task 1 for Part 2 of the survey.....	112
Figure 6-1: Research Design of the SOVAT evaluation study.....	121
Figure 6-2: Rate of completion (between groups) using SPSS during the pilot study. ....	123
Figure 6-3: Mean time per task per period for SOVAT and SPSS-GIS.....	128
Figure 6-4: Satisfaction scores by period (A lower number is better). ....	132
Figure 6-5: User preference between SOVAT and SPSS-GIS.....	133

# **1. INTRODUCTION**

## **1.1. Introduction**

The development of numerical-spatial routines is frequently required to solve complex problems. Individuals who use decision support technology need a system that is capable of supporting the development of numerical-spatial routines. Currently, there is no decision support system that is effectively able to accomplish this task and thus decision makers are forced to improvise these steps with individual software applications or manual labor-intensive processes.

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 components (or steps) the individual uses for solving a given problem. There will be two types of components highlighted: spatial components and numerical components. A numerical component only involves numerical data, while a spatial component involves spatially-defined data (for example objects that have a coordinate value). For the purposes of this dissertation, a numerical-spatial routine (or numerical-spatial problem solving) is one that contains both numerical and spatial components and attempts to solve numerical-spatial problems. As an example, a typical routine during community health assessment involves both spatial and numerical components, and can be described as such [1]:

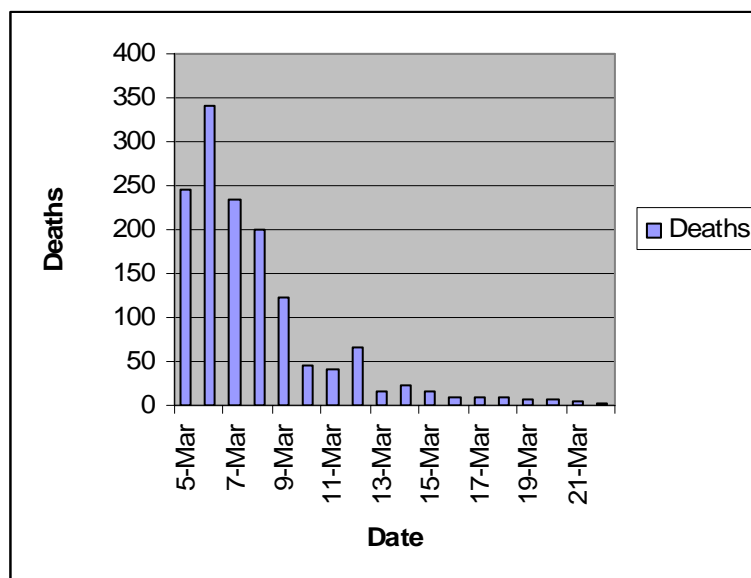
1. Identify geographic community of interest
2. Identify health factors within the community
3. Identify bordering communities of interest
4. Identify health factors within bordering communities

5. Compare factors within community against factors of bordering community
6. Identify aggregate (state-wide, or national, etc) community
7. Identify health factors within aggregate community
8. Compare factors within community against factors of aggregate community

The first step of identification of a geographic community is a spatial component. 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 5 diseases per 100,000 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 may decide to first 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 serves as the spatial component, while the calculation of disease ranking serves as the numerical component.

A more in-depth example of numerical-spatial problem solving can be seen by looking as far back with the work of John Snow. Snow, who is known as the “father of epidemiology”, eradicated the Cholera outbreak in London during the middle of the 19<sup>th</sup> Century. He did this by

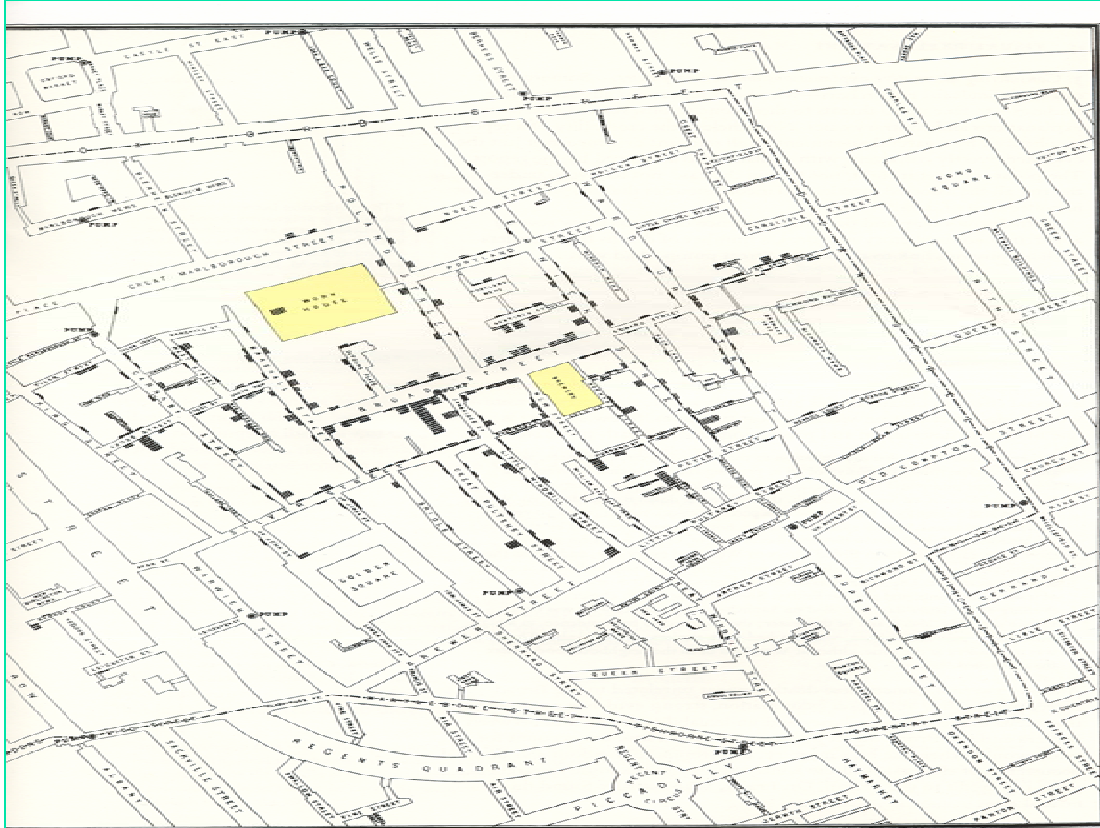
developing a numerical-spatial routine [2]. He first realized that a numerical-only routine used to answer the question “*How many deaths are from Cholera by day from the start of the epidemic*” was not sufficient in solving the puzzle. This can be seen via bar chart in Figure 1-1 (adapted from [2], page 29). This is because numerical problem solving for a identifying the cause of the outbreak was not sufficient; “descriptive narration is not casual explanation” (page 29, [2]). Snow realized the importance of combining numerical and spatial components in order to solve this dilemma. He did this by creating a map of the streets of London (with houses, bars, water pumps, and other important city markers). He then mapped each death (individual’s residence) to a spatial location on the map (Figure 1-2<sup>1</sup>). This is the same as using GIS to geo-code data points on a digital map (GIS will be discussed in greater detail in section 1.6).



**Figure 1-1: Results showing descriptive counts of cholera deaths.**

---

<sup>1</sup> Reprinted from Tufte E, Visual Explanations: Images and Quantities, Evidence and Narrative, pg. 30-31, 1997, with permission from Graphics Press, Cheshire, CT



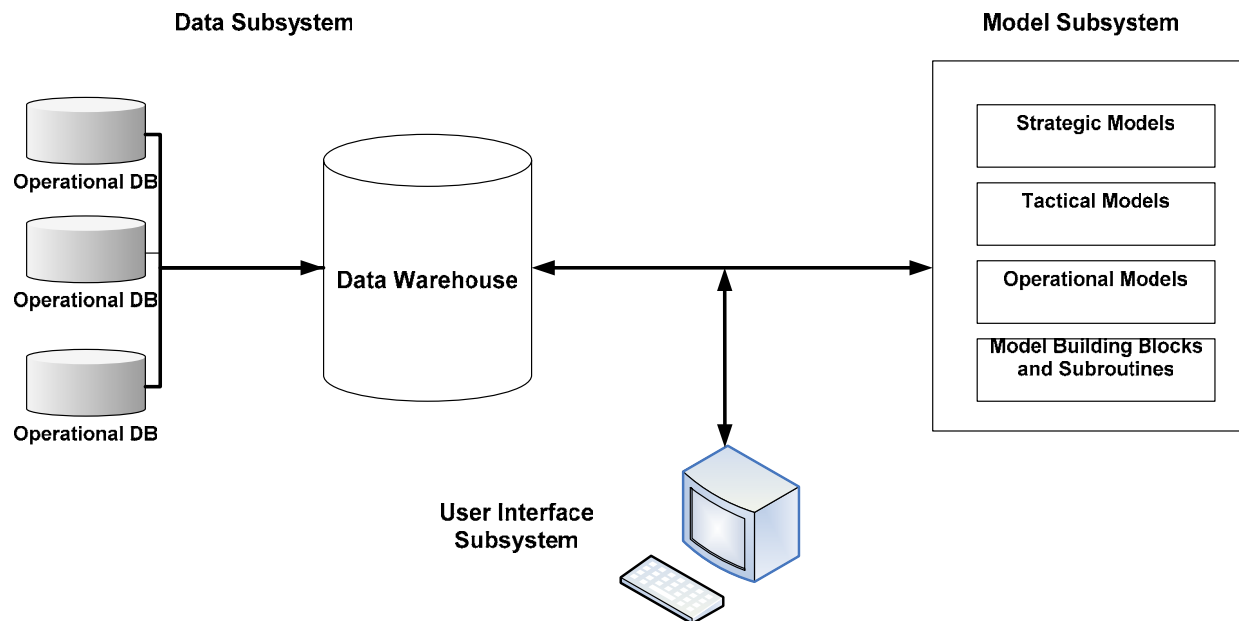
**Figure 1-2: Numerical-spatial problem solving**

Snow created a numerical-spatial routine and determined the source of the outbreak without the help of a decision support system. This method of solving spatial and numerical problems can be implemented with today's technology including Geospatial Information Systems (GIS), On-Line Analytical Processing (OLAP) and statistical software packages. OLAP and GIS will be addressed in this chapter (sections 1.5 and 1.6 respectively) and will be thoroughly examined in the next chapter.

## **1.2. Decision Support Systems (DSS)**

Decision Support Systems (DSS) were originally developed in the early 1970's and were first defined by Gorry and Scott-Morton as systems that support unstructured decision making [3].

Over the years, many different frameworks and models have been proposed to signify decision support technology. Simply stated, a ‘typical’ DSS has a user interface subsystem, a model (or problem processing) subsystem, and a data subsystem. Different terms have been used in different papers, but the fundamental description of these components has been the same. An illustration of this can be seen in Figure 1-3 (adapted from [4], p. 64).



**Figure 1-3: An example of a traditional DSS architecture.**

In addition, development of a DSS architecture consists of three-tiers:

1. DSS Tools –Constitute the low-level programming languages used to construct the software.
2. DSS Generators – The middle components that represent the unique software included in the system and are built by the DSS tools.
3. Specific DSS – The actual system that is used by the decision maker which is a culmination of all the DSS generators.

For the most part, the data subsystem offers nothing unique to decision support system research except for extraction [5] of the raw data into an aggregated form suitable for high-level decision making. Due to this, not many of the DSS papers have focused exclusively on this component. The modeling subsystem, in contrast, received significant attention in the literature and was recognized as the integral component within the architecture [6]. A ‘model’ is a representation of a real-world problem that is often organized by a set of mathematical subroutines containing decision variables, uncontrollable variables, and outcome variables [7]. Most of the modeling details in these early systems were hidden to the user. The learning curves for model construction were sometime high which resulted in significant difficulties with interaction and modification of the model; the user had to be knowledgeable in the use of different models, and know the specific parameters for instantiation of these models [8]. A user in constructing a model had to understand the mathematical relationships between these different types of variables. For example a mathematical model for a financial investment company would consist of decision variables (investment alternatives and amounts, how long to invest, when to invest), uncontrollable variables (inflation rate, interest rate), and outcome variables (total profit, risk, return on investment, earnings per share, etc) [7].

The user interface subsystem was another area of focus with these early systems in the 1970s and 1980s. The user interface subsystem was examined in relation to designing interactive command-line interfaces that supported English-like querying. With the early DSS, the entire human-computer interaction existed through a command-line interface which added complexity by requiring knowledge of specific command-line syntax and rules. Keen [9] outlines the difficulties with these early command-line interfaces for query construction. He uses a linear programming model for a financial system as the example.

As the field of Artificial Intelligence (AI) began to interact with DSS, many papers highlighted Expert Decision Support Systems [8, 10-23] that attempted to strengthen the model subsystem component. Many of these mentioned the area of intelligent model selection [12, 15, 23]. While this did provide the appearance of enhancing DSS capabilities, it wasn't until the development of User Interface Management Systems (UIMS) that the link between user interface and model subsystem strengthened and the focus shifted from a data (and model-oriented DSS) to a user-centered DSS [12]. Since the 1990's, the theme among most designers was that the user interface subsystem was the most crucial factor in DSS design rather than the problem-processing or model subsystem that constituted the earlier systems. Systems that offered visual development of queries gave the user a non-complex, easy-to-use graphical interface with which to comprise and analyze numerical queries. Instead of worrying about command-line syntax, users were now able to construct routines through mouse clicks on Graphical User Interfaces (GUIs). Queries performed with command-line interfaces could now be developed using simple data entry forms and mouse clicks. This ability made the decision support systems much less intimidating and more appropriate for novice users and brought decision support systems into the mainstream market.

### **1.3. Problem Solving Routines by Type of DSS**

Decision Support Systems can be classified by the type of support they offer, and these types can dictate how problem solving routines can be developed. The two categories to consider are either model-oriented or data-oriented. A model-oriented DSS places much more emphasis on the use of mathematical models to solve 'if-then' problems. In this context, routine components are generally thought of as scenarios because the user is concerned with examining future

possibilities rather than historical outcomes. Scenarios typically involve graphical iconic representations in a discrete-event or continuous simulation format. Here the real-world scenario is represented by a series of nodes and icons with arcs that connect them together. Each node represents a point in the scenario lifecycle process. The iteration through the scenario is dictated by mathematical distributions and processes. A data-oriented DSS is more focused on the retrieval of historical information from the data subsystem component. Here, end-user functions are more important than model construction and simulation of futuristic possibilities. Tools that provide for efficient and powerful analysis and display of results are critical for decision support. Graphical user interfaces are only one of the necessary ingredients for development of routines with a data-oriented DSS. Both a powerful data subsystem and a user interface subsystem are critical for this type of research.

#### **1.4. Issues with Data-oriented Decision Support for Development of Routines**

Relational database technology is not sufficient for ad-hoc numerical or spatial problem solving development. Traditional DSS use relational database generators (the middle components that represent the unique software included in the system and are built by the DSS tools) to store the numerical data. GIS, which deals with spatial data, has also been built to support this flat-file format. Thus the transactional and processing time for both numerical and spatial functions is very slow due to the operational and complex joins that must occur. This slows down analysis and impedes decision making. Even a seemingly simple query such as *“Find the fisherman who earn less than \$30,000 and catch swordfish”* can be difficult for a transactional system to perform. There are data warehouses that extend numerical data beyond this transactional format to a more multidimensional model. However, the difficulty lies in extending the spatial

information so that it can be coupled with multidimensional numerical data but still be accessible to the GIS which is based on the transactional data model.

The data subsystem is not the only area of concern for development of effective numerical-spatial DSS. Developers must create a proper UI subsystem that realizes the synergy between the numerical and spatial data. Users must be able to view these different types of information on the screen in the proper spatial and numerical format. This will potentially allow users to create numerical-spatial routines and realize the complete integration between numerical and spatial entities within the DSS.

DSS generators are important aspects in data-oriented decision support systems because they drive the data model and the nature of the retrieval, analysis, and presentation of the information taken from the data model. Generators have the potential to be specific DSS but are lacking in at least one of the three subsystem areas. Many community health decision makers use generators for their analysis, such as GIS (such as ArcView) and statistical software (such as SAS). Enhancement to these technologies can shift them from generators to Specific DSS. Since much of the data in decision making entails some form of spatially-related information, decision makers need the proper DSS generators that allow them to perform spatial functions. The development of spatial routines could be significantly enhanced by utilizing a more powerful data subsystem component such as On-Line Analytical Processing (OLAP). OLAP stores its data in a multidimensional cube format and the data is pre-aggregated and fully materialized. Thus, with OLAP, there is no need for performing joins during querying. The data is already summarized and thus only needs to be fetched. In addition, data exploration is significantly facilitated and enhanced due to the unique multidimensional data model. Data exploration is a significant part in development of both numerical and spatial routines as the user searches for the

appropriate variables to include. This might require examining data at a lower level of granularity or looking at data from a different angle; both of which can easily be performed using OLAP. GIS is essential for developing spatial routines. There is no alternative technology that could simulate this process. Therefore it must be coupled with a DSS generator that can, by itself, support development of powerful numerical routines. In addition this generator must extend GIS beyond a two-dimensional format for supporting spatial functions. In addition, it must enable for the development of an interface that supports not only the combination of numerical and spatial information display, but also an interactive easy-to-use environment for creating different types of numerical and spatial routines. The potential of this synergy is a system that can significantly enhance spatial and numerical problem solving.

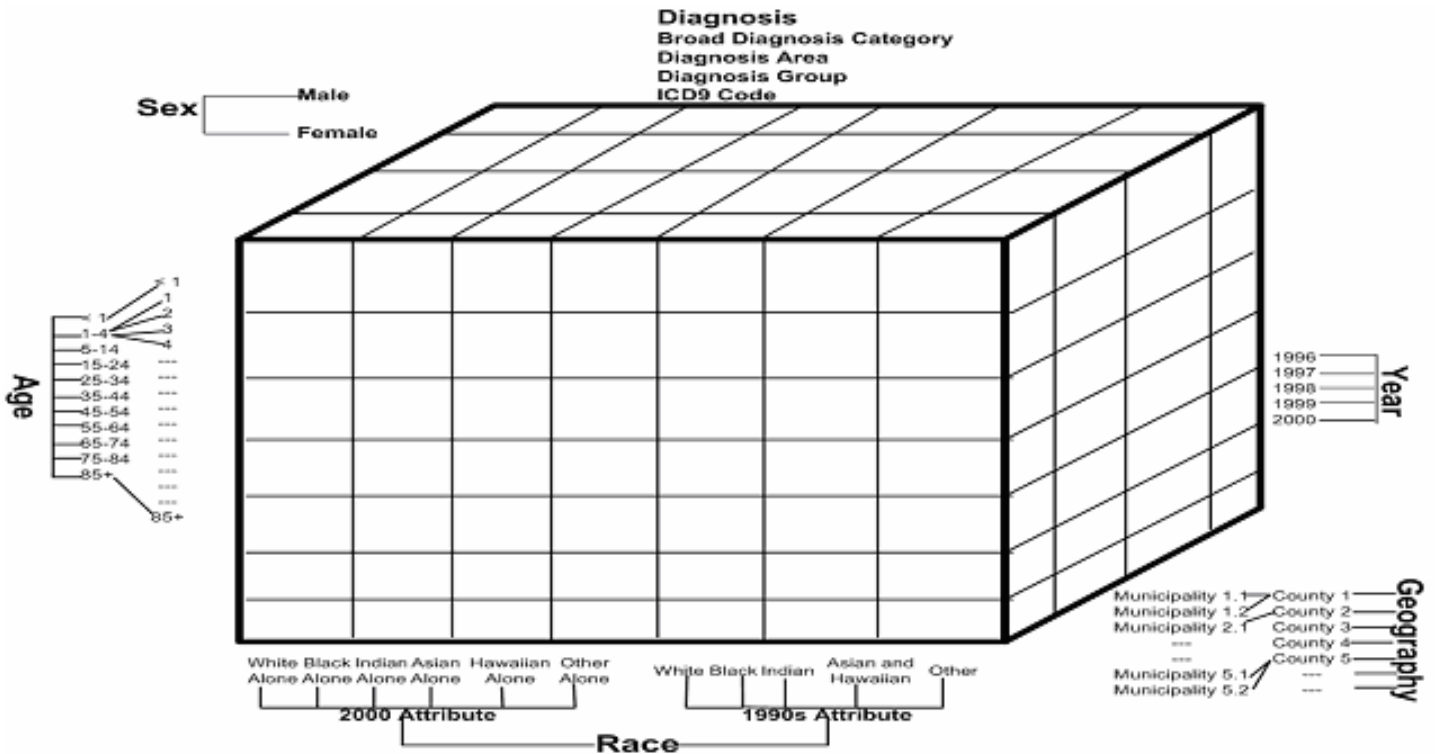
Despite these advantages, OLAP has not been utilized for development of combined spatial and numerical routines. A brief background of this technology will now be discussed, followed by an introduction to GIS

### **1.5. On-Line Analytic Processing**

On-Line Analytical Processing (OLAP) is a data warehousing technology that supports the creation of a multidimensional data model. The data is conceptually modeled as a multidimensional cube (Figure 1-4)<sup>2</sup>. An OLAP cube contains dimensions and attributes that define the dimension. A cube must contain data on at least one numerical measure. The combination of attributes in a cube's cell forms the numerical measure value for that cell.

---

<sup>2</sup> Reprinted from Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05), Scotch M Parmanto B, SOVAT: Spatial OLAP Visualization and Analysis Tool, 2005, pg. 142b, permission from IEEE Computer Society.



**Figure 1-4: A Multidimensional OLAP Cube.**

OLAP supports very fast and efficient ad-hoc querying because the data in the cube is pre-aggregated and can be fully materialized. There are four functions that are distinct to OLAP for data analysis:

- Drill-up – go from a low level to a higher level of data (ex. month → year).
- Drill-down – go from a high level to a lower level of data (ex. year → month).
- Slice – At a particular level, examine data for a particular attribute (ex. At the year level examine 1999).
- Dice - At a particular level, examine data for particular attributes (ex. At the year level examine 1999 and 2000 and at the City level examine Toronto and Vancouver).

Pivot is another OLAP function, which switches the rows and columns as would be done in a pivot table. Using OLAP, the user can quickly analyze data in an ad-hoc manner by drilling up

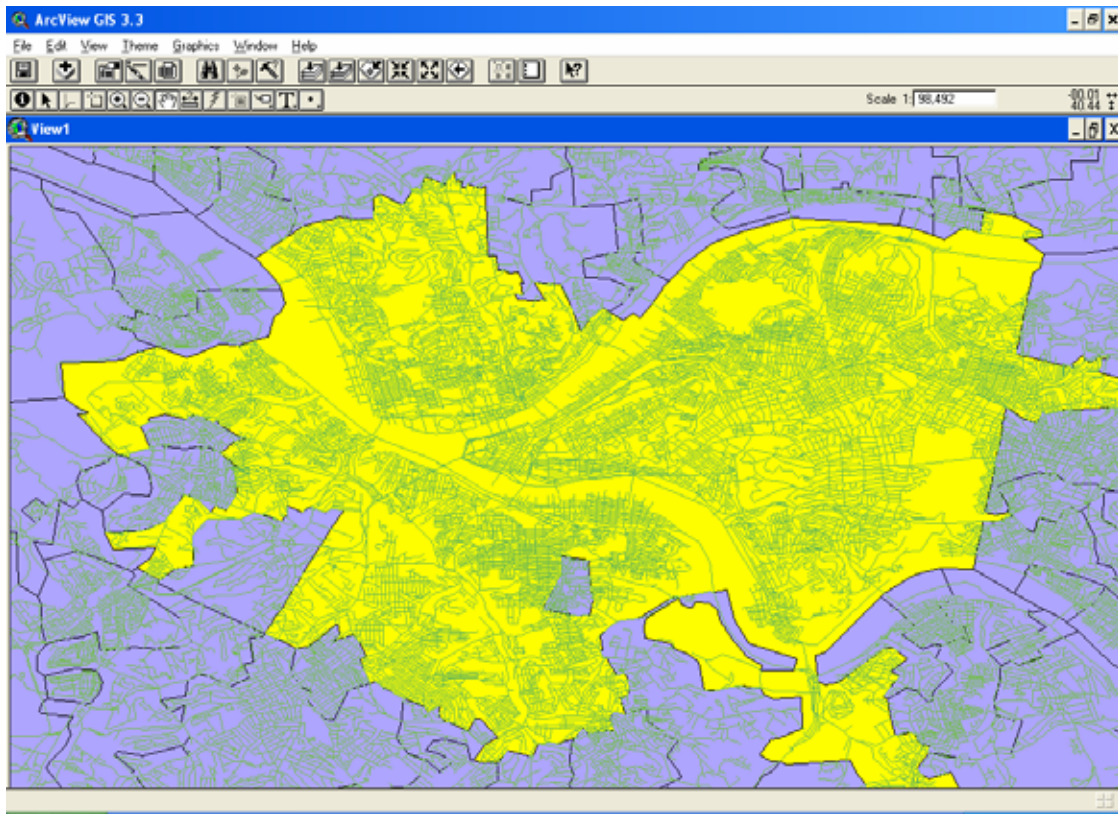
or down, and slicing and dicing on dimensions within the cube. Data retrieval occurs within seconds as compared to a traditional on-line transactional processing system (OLTP) which has the potential to take several minutes to return the results. OLAP has recently been recognized as an important tool for decision support, yet is still not as popular as traditional database systems.

## **1.6. Geospatial Information Systems**

Geospatial Information Systems (GIS) have been around for many years as a tool for projects such as city planning, assigning environment protection areas, and telecommunications research. It is recognized as a technology for analyzing, displaying, and manipulating spatial data. A GIS environment consists of layers (land, rivers, roads, buildings, cities, etc) on top of one another to form a detailed digital map. GIS technology is not intended to just display spatial information; it supports several powerful analysis functions such as determining best routes between two locations (network analysis), buffering (specified distance around a particular location), and geocoding (mapping coordinate points on a map). GIS has gained popularity in the realm of decision support due to the abundance of spatial information and the growth of familiarity of GIS technology, through applications such as ArcGIS<sup>3</sup> (Figure 1-5).

---

<sup>3</sup> <http://www.esri.com> (11/05/04)



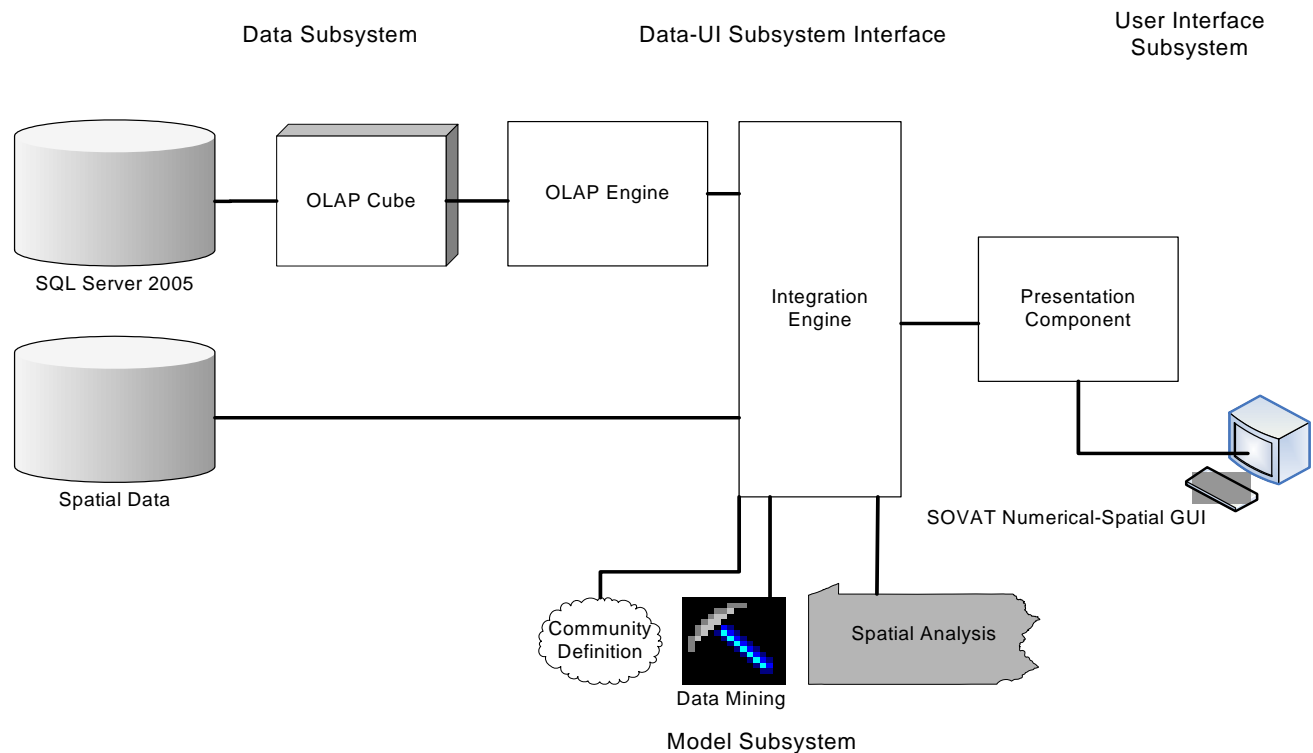
**Figure 1-5: An example of GIS application (here using ArcGIS).**

As previously mentioned, the issue with GIS in relation to DSS use is that numerical and spatial query development is limited to the transactional backend architecture that these systems frequently support. Thus ad-hoc interactivity with the underlying data is rigid and two-dimensional. There is no ability to drill down on dimensions or slice and dice for specific attributes in order to create preferred spatial and numerical queries. These processes can be mimicked in the 2-dimensional environment that GIS supports, but the efficiency and query development time for the user is lengthened.

## **1.7. Research Focus**

This dissertation will examine whether combining On-Line Analytical Processing (OLAP) and Geospatial Information System (GIS) technology creates a more useful data-oriented decision support system for numerical and spatial problem solving. For the purposes of this dissertation,

the term “OLAP-GIS” will be used to describe a system that combines these two technologies. The Spatial OLAP Visualization and Analysis Tool (SOVAT) has been developed in this framework (Figure 1-6). The SOVAT architecture extends beyond the traditional decision support system model (Figure 1-3<sup>4</sup>) to combine numerical and spatial components (this architecture will be discussed in more detail in section 3.2).



**Figure 1-6: SOVAT architecture.**

A case study in the domain of community health assessment will be used to evaluate the system. The complete evaluation will contain three studies:

1. Usability Study – Aimed at identifying human-computer interaction (HCI) issues with an OLAP-GIS system.
2. Community Health Survey – Aimed at identifying DSS technology used by today’s CHA researchers.

<sup>4</sup> Reprinted from International Journal of Medical Informatics, In Press, Scotch M Parmanto B, Development of SOVAT: A numerical-spatial decision support system for community health assessment research, 2005, with permission from Elsevier

3. Summative Evaluation – Aimed at comparing an OLAP-GIS system against current information technology in relation to numerical-spatial problem solving.

These studies will be described in more detail in chapters 4, 5, and 6 respectively.

## 2. BACKGROUND

### 2.1. Overview

This background chapter includes two sections related to the current work. The first section contains common DSS generators (or applications) used in decision support systems for either spatial or numerical problem solving and shows their contributions and limitations in these areas. The second section examines some spatially-related decision support systems and examines their success and limitations at supporting numerical-spatial routines.

In order to effectively examine numerical-spatial problem solving, it is necessary to look at what features are important for a numerical-spatial DSS to contain. After considering these features, the technologies that support them will then be examined.

Users of decision support systems are often required to ‘dig through’ mountains of data in order to *discover interesting patterns* that will lead to *knowledge discovery*. This allows them to make changes to current systematic processes based on these discovered patterns. Simple query development using descriptive functions is incapable of obtaining these results. Special functions under the practice of data mining must be used in order to discover interesting patterns within the data.

The mountains of data that decision makers have access to must be stored in a data warehouse that allows for the storage of *large and complex data sets*. Traditional relational databases are not sufficient for the purposes of high-level problem solving. Technology must be used that aggregates complex data into a form suitable for decision support while maintaining data integrity.

Decision makers also need a system that enables for both *multidimensional viewing* and multidimensional detailed data exploration. Efficient navigation is very difficult using a transactional database. Viewing the data in a multidimensional tree-like fashion displays the hierarchical relationship of the underlying data. Being able to efficiently explore different granularities of the data facilitates query development and enhances problem solving.

The importance for the discovery of interesting patterns using data mining functions was already discussed. *Statistical analysis* is similar in the respect that it goes beyond simple descriptive analysis such as counts, sums, min, and max that provide only a partial story of the data. Inferential statistical analysis is essential for decision support and is found in many DSS.

Decision makers using spatial data need to *view information* in a spatial format. In some cases, paper maps are sufficient. However, the abundance and growth of spatial information has heightened the need for digital presentation. This requires a system to display spatial information in a digital format. In addition, spatial functions can not be performed using paper maps. Decision makers need a system that can perform different types of *spatial analysis* and then plot this information on the digital map.

As mentioned, CHA requires a system that can perform *numerical-spatial problem solving*. Such a situation was shown with the John Snow example. Combining numerical and spatial components requires many of the features previously listed. Hence, it is difficult to conduct numerical-spatial problem solving with a DSS that does not contain the before-mentioned features. That being said, the mere existence of these features does not guarantee that the system will easily be able to combine and integrate spatial and numerical components. A solution to this issue will be discussed later in section 3.2.1.

The features previously mentioned: interesting patterns/knowledge discovery, complex data sets, multidimensional viewing, statistical analysis, viewing of data (spatial and numerical), spatial analysis, and numerical-spatial problem solving, are all essential for an effective numerical-spatial decision support system. A review of the literature will show how there are only a few complete systems that combine the necessary technologies (or DSS generators) that support these features. After considering each technology, it is anticipated that a combination of On-Line Analytical Processing (OLAP) and Geospatial Information Systems (GIS) is the best solution for addressing this problem.

## **2.2. Technologies for Supporting Spatial and Numerical Problem Solving in DSS**

Table 2-1 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 previously addressed. As evident, most of these technologies address some, but not all of the essential components for numerical-spatial problem solving. There are a few reasons for why so few comprehensive systems 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. Results generally need to only be displayed in tabular-numerical form. Because of this, many software packages have unimpressive visual interfaces. 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 in coupling spatial functions, such as geo-coding or color gradation of spatial objects within a coordinate system.

**Table 2-1: DSS generators for decision support and the capabilities they provide.**

	Interesting Patterns/ Knowledge Discovery	Large, Complex Data Sets	Multidimensional View/ Navigation	Statistical Analysis	Spatial Presentation	Visual Charts	Spatial Analysis	Numerical- Spatial Problem Solving
Statistical Software				X		X		
Data Mining	X			X		X		
GIS Software					X		X	
Relational DB		X						
OLAP		X	X			X		

These technologies: statistical software, data mining, GIS, relational databases, and OLAP, will be addressed individually, with emphasis on their strengths and weaknesses for numerical-spatial decision support.

### **2.2.1. Statistical Software**

Statistical functions are considered one of the most important characteristics in a decision support system. The functions act on information from the data subsystem (inputs) and produce data (output) for end-user analysis. Statistical functions come in many forms within DSS architecture. With a model-oriented DSS, they are managed within the problem-processing subsystem with input from the data subsystem. Statistical functions are frequently located within models, whether it is optimizing, non-optimizing, or simulation models. Many times statistical software packages such as SAS<sup>5</sup> or SPSS<sup>6</sup> are embedded into DSS architecture within the problem-processing subsystem. For example, looking at Figure 1-3, the statistical generator

<sup>5</sup> <http://www.sas.com/> (11/01/04)

<sup>6</sup> <http://www.spss.com/> (11/01/04)

would most likely be located in the “building blocks and subroutines” section of the model subsystem and be able to support any of the models above it.

As their analysis capabilities have become more powerful, more people have considered stand-alone statistical packages to be specific DSS. This is not the case. While they perform well with problem-processing and modeling, they typically lack sufficient language and/or data subsystem capabilities necessary for an effective DSS. For example, many statistical software packages have poorly designed user interfaces. In addition, the backend components of many of these software packages are rigid and two-dimensional. Improvements have been made in these two areas, but they are still considered DSS generators rather than a full fledged specific DSS. Some have termed these Statistical Decision Support Systems (SDSS) defined as “information systems whose purpose is to answer queries that are requests for statistics about an underlying set of raw data when those answers are going to be used as an input to some decision” (p. 375, [24]). It is noted that these “decision support systems” are similar to statistical databases. Data-oriented decision support systems, that do not place heavy emphasis on models, incorporate statistical functions within the database management system.

A decision support system whether it is model-oriented or data-oriented, must contain numerous types of statistical functions for effective data analysis. Use of off-the-shelf statistical packages can be effective DSS generators within a specific DSS. These systems provide a wide array of statistical functions from mean and median to multiple regression analysis.

Decision support systems aimed at solving numerical-spatial problems need statistical functions that factor in spatial proximity. For example, in order to account for sparse geographic areas, smoothing functions will factor in data from nearby larger geographic regions. This process reduces data variability and allows numerical patterns to be discovered [25]. Spatial

smoothing is best used when decision makers need to analyze clusters of geographic regions (it does create bias in estimates when examining individual small areas). For example, a community health assessment expert needing to analyze health trends across multiple counties would most likely use spatial smoothing to account for sparse areas of population. Thomas and Carlin [26] used spatial smoothing to identify late detection areas of breast and colorectal cancer in Minnesota counties. Spatial smoothing was important for Thomas and Carlin's study because there are variations of high populated and low populated counties within the state; many of which border each other. In order to get an accurate assessment of areas with late detection rates, spatial smoothing needed to be performed.

Many GIS packages contain plug-ins or extensions that perform spatial smoothing. For example, ESRI's ArcGIS offers two plug-ins, *Spatial Analyst* and *Geo-statistical Analyst*, which perform a spatial smoothing function called "kriging". Other techniques involve coupling the GIS generator with the statistical generator. This allows the spatial smoothing to be calculated within the statistical model with the results passed to the GIS for display.

### **2.2.2. OLAP**

While transactional databases are designed to capture data, data warehouses are designed to extract information out of the database. This fundamental difference suggests that the data warehouse should be designed according to a set of different principles from the traditional transactional database. The set of design principles for a data warehouse is called multidimensional modeling, often referred to as the star schema approach. Dimensional modeling is a logical design that seeks to present the data in a standard, intuitive framework that allows high performance access [27]. On-Line Analytical Processing (OLAP) supports multidimensional data warehouse modeling. OLAP allows for rapid queries of multidimensional

data that enables for powerful analysis and research discoveries through display on easy-to-use front-end interfaces. Through its client/server architecture, OLAP technology enables the user to view different dimensions of multiple datasets and then query several dimensions at once in order to view their relations to one another. OLAP is an ideal model for data warehouses and decision support systems with its quick retrieval of pre-aggregated data.

In the realm of decision support system architecture, OLAP represents the data subsystem replacing the traditionally used relational schema. The data subsystem by itself has received little attention in decision support system research. This is because database management is not inherently different in decision support system architecture than in an MIS, OLTP systems, or EIS. The one difference is in relation to data extraction, where daily transactional data is automatically pulled, pre-processed, and aggregated into a format suitable for high-level decision support. This unique DSS process has been highlighted by Methlie [5] in relation to GADS (an early Spatial DSS). Numerical queries can be developed using OLAP by browsing the multidimensional data with the use of unique (drill-down/up, roll-down/up, slice and dice) functions. OLAP also can be used with elaborate front-end visual clients that display the results of the numerical query.

One of the first papers to introduce OLAP within a decision support framework was by Koutsoukis et al [28]. The author specifies the use of OLAP functions such as drill-down, drill-up, slice/dice, as pivotal elements for this process. He describes using OLAP functions to initially produce an LP-optimization model using DOME (Domain Modeling Environment) [29] software developed by Honeywell to support numerical query development. In their example, the data is coupled to the model, which allows the user to create different queries of aggregated

and disaggregated models, thus the level of data (defined by Roll-up/Drill-down OLAP functions) drives the level of the query detail.

Research examining OLAP and decision support has mainly been focused around data modeling rather than decision modeling. The multidimensional representation of the data in OLAP architecture provides several advantages for storage of data for decision support, yet the unique functions and easy-to-use presentation layer software it supports have not been significantly used for decision modeling and query development.

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 DSS generators 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 [30] as well as the lack of realization of the potential of this combination. At the current time, there are only a few OLAP-GIS systems used for decision support. These will be discussed later on in detail in section 2.3.

### **2.2.3. GIS**

GIS is not considered a decision support system based on the fact that it lacks the ability for supporting problem specific modeling [30]. In the realm of the DSS software development process, it is typically considered a DSS generator rather than a specific DSS. 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. To the user, the environment for query

development within a spatial context *is* the GIS interface. This is the area on the screen where routes and layers can be created, objects such as buildings and roads can be manipulated, and analysis can be conducted. In the realm of the DSS components, the GIS interface constitutes the user interface subsystem. 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 DSS generators should be incorporated with GIS in order to create an effective support system for creating spatial and non-spatial queries. None of the previous decision support systems have adequately addressed this problem. One of the earliest decision support systems to incorporate a geospatial component was the GADS system [31]. The system which was developed in the '70s had limited geospatial functionality, but is still recognized as a pioneer within Spatial Decision Support System (SDSS) literature. The Tolomeo system developed in the 1990's signified great strides in the realm of interactively modeling spatial queries, but still is limited in the area of query development since it is only a DSS generator rather than a specific DSS [12]. Thus, the system was an ineffective tool for combining spatial and non-spatial tasks. In order to overcome many of the inflexibilities that GIS stand-alones have in relation to problem specific modeling, GIS needs to be incorporated with other technologies and DSS generators such as Online Analytical Processing. GIS provides many useful functions for spatial problem solving such as buffering, network analysis, overlaying, and clipping. Buffering is frequently used in disaster planning, service/coverage area identification, and conservation area assignment. Buffering even has even been used recently in the war in Iraq for military installation force protection planning [32]. Network analysis is useful for determining shortest path and best route problems. They are

highlighted in many online GIS sites such as MapQuest [33], and used for private purposes such as hydraulic network modeling.

In addition more accurate spatial analysis can occur through the implementation of statistical methods within GIS. For example, spatial smoothing and empirical bayes smoothing calculations can be implemented in order to address the *small number problem* [34].

#### **2.2.4. Data Mining/Knowledge Discovery**

More recent analysis efforts in DSS have been to add data mining functions such as clustering, decision trees, neural networks, association rules, and Bayesian networks as techniques to perform classification and prediction tasks. This research area called Knowledge Discovery in Databases (KDD) has become popular in areas such as insurance, direct-mail marketing, telecommunications, retail, and health care in order to find useful patterns within the data for decision support [35].

On-Line Analytical Mining (OLAM) is the term referred to the integration of data mining functions within an OLAP environment. This synergy is unique to data mining since most data mining DSS generators are designed for traditional transactional databases. Due to this fact, there are a certain set of data mining functions that are appropriate for OLAM, and a certain set that are better suited for the traditional database approach (Table 2-2, from Parmanto, unpublished). The multidimensional data model and the unique exploratory functions of OLAP are much better suited for the data mining methods such as: association rules, decision trees, Bayesian Networks, and clustering. OLAP allows for traditional techniques such as association rules to be enhanced by allowing for multi-level, multi-dimensional association rules to be generated. For this reason, Microsoft's SQL Server 2000<sup>7</sup> only allows for clustering and decision tree techniques to be used for creation of multidimensional mining models. The techniques such as neural networks and

---

<sup>7</sup> <http://www.microsoft.com/sql/> (11/01/04)

genetic algorithms that require more detailed-level data are better suited for an offline environment.

**Table 2-2: Data mining functions for OLAM.**

<b>Data mining models</b>	<b>Data mining methods</b>
Online data mining (OLAM)	Association Rules, Decision Tree, Learning Bayesian Network, Clustering
Offline data mining	Neural Networks, Genetic Algorithms, K-nearest Neighbor, Instance-based learning

The needs for mining during numerical-spatial problem solving are different than the needs for traditional problem solving (that considers purely numerical issues or purely spatial issues). For example, many bio-surveillance experts would like to group cases of diseases together by both space and time factors. This allows them to gauge whether there is an outbreak of a certain disease. One manner in which this can be performed is through spatial clustering. This technique [36-46] has been very popular in spatial decision support systems. However the techniques for performing spatial clustering such as genetic algorithms and K-nearest neighbor are more suited for offline transactional systems. Thus using OLAM to perform these tasks is not appropriate. This technique is much different than numerical clustering of spatial objects that is popular in decision support systems that deal with both numerical and spatial unstructured problems. The difference is that the spatial objects (counties, states, etc) contain numerical attributes (cancer deaths = 'X') and the classification is based on these attributes and not the spatial relationship from one spatial object to another. Thus for numerical clustering, numerical relationships are much more important than spatial relationships (there is no concern about distance). On the contrary, spatial clustering techniques use proximity between objects to form clusters. For example, clustering counties based on cancer rate is numerical clustering, while clustering counties based on distance to tertiary hospitals is spatial clustering.

One of the defining techniques in spatial clustering is constraint-based cluster analysis such as clustering with obstructed distance (cod). In this instance, physical obstacles that exist in nature such as mountains and rivers are factored into the cluster analysis [47]. For instance, a bank planner looking to add additional ATMs would factor in local rivers in deciding the best possible locations of the new machines. Thus, he/she would probably not add all the machines to one side of the river since it would be difficult for the customers on the other side to access them [47]. This type of issue is not a concern with traditional clustering techniques.

Simple numerical clustering is different in numerical-spatial DSS because even after the calculations are performed, there needs to be a coupling between other DSS generators. For example, the results need to be sent to the GIS software so the clusters can be appropriately represented on the digital map. Many off-the-shelf data mining software applications do not place emphasis on visual display of classification techniques. In order to solve combined numerical and spatial issues, this is a very important criterion, since the user needs to see which spatial objects are classified together. This involves combining GIS and OLAM by displaying the results from the mining algorithm on the digital map in a color coordinated fashion that matches spatial objects to their appropriate cluster. This methodology is drastically different than using color gradients on a choropleth map to identify numerical differences of geographical areas based on a particular measure (such as disease rate). Color gradation is a GIS function, where numerical clustering is a data mining/knowledge discovery function.

Manhinhakumar et al [48] provide an example of multivariate nonhierarchical numerical clustering involving spatial objects based on environmental data. The authors hoped to identify ‘ecoregions’ or regions of ecological similarity. They were able to develop clusters of 100, 500, 2000, and 3000 ecoregions. Multi-level, multidimensional association rule generation is another

technique that could be useful for classification within OLAM for spatial decision support. This type of research would be very valuable to community/public health assessment experts using a decision support system. Despite this fact, numerical classification of spatial objects is rarely used for this type of research.

### **2.3. Decision Support Systems for Numerical-Spatial Problem Solving**

Gao et al [49] describes a framework for numerical-spatial problem solving for Spatial Decision Support Systems (SDSS). This article highlights the inadequacies in problem solving within a spatial context. The author mentions how there has not been a successful framework for addressing Spatial Decision Making (SDM) while most users feel that the process involved in solving complex spatial problems as being unsatisfactory. The authors provide examples of numerical-spatial routines (page 7 of [49]). The intersection of outputs between spatial and non-spatial components ( $\text{component 1} + \text{component 2} = \text{component 3}$ ) represents a problem solving routine which can address numerical-spatial problems. While the authors claim that the creation of numerical and spatial components creates for a flexible SDSS architecture, they admit the difficulty in creating complex numerical-spatial routines with a single step (only one component) [49]. Their solution is to create spatial and numerical components independently and then combine them into a complex model as the one previously described. This approach makes some very important strides in numerical-spatial problem solving. Further research needs to be done by looking at how specific technologies can be added to enhance this process. For example, combining On-Line Analytical Processing (OLAP) and GIS can facilitate multidimensional numerical and spatial problem solving. As will be described later, combining

OLAP and GIS addresses this problem and provides the capability of creating numerical-spatial routines within a single step.

Bedard et al [50] is one of the few individuals who has combined OLAP and GIS into a decision support system. His work represents the only publicly deployed OLAP-GIS system called JMAP Spatial OLAP [51]. A prototype of this work is the ICEM-SE project. The authors developed the ICEM-SE project to integrate geographic knowledge discovery for Canadian environmental health surveillance. The user interface of ICEM-SE was developed using off-the-shelf plug-in technology provided by ProClarity<sup>8</sup>, 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 (although this is never stated in [50]). This data is then used to process the OLAP cube. For a detailed look at the ICEM-SE architecture, please see [50] (page 92). 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

These abilities, which are only available through the combination of OLAP and GIS, provide the user with a powerful environment for numerical-spatial problem solving. Incorporating these two DSS generators into a specific DSS can produce a useful research tools

---

<sup>8</sup> <http://www.proclarity.com> (11/18/04)

as seen on pages 83-85 of [50]. 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 specific dimensions in order to switch from one level of aggregation to another. The result is then displayed on the map. While ICEM-SE system has many positive traits for numerical and spatial problem solving, there are some areas where it is lacking. The most significant issue that ICEM-SE has is the lack of manipulation of spatial objects directly on the map. While specific geographic regions can be highlighted by mouse click, the actual drill down operation does not occur by the map, but rather through the selection of OLAP functions on the sidebar. In addition, OLAP-like functions such as “drill-out” are non existent. This limits the effectiveness of numerical-spatial problem solving.

While the user has the ability to use the OLAP dimension trees to create a query, the same can not be said for the use of the map. There is no ability to drill-down, drill-up, slice, dice, or drill-out on the geography dimension through the map itself. This gives the impression that the map is more of a visualization tool, rather than a powerful DSS generator. Also, it is not clear what spatial functions (if any) are available. If the system is to be used as an environmental health decision support system (as it is claimed), it would seem logical to incorporate spatial analysis functions such as buffering for tasks such as environmental disaster assessment (without it would seemingly lessen the potential of the OLAP-GIS synergy).

Another area where this system is lacking is in the displaying of results. Through the use of OLAP and GIS, the developers have combined spatial and numerical data into a decision support

system; however, they do not provide simultaneous results with both spatial and numerical representation. For example, on page 85 of [50], the author shows a spatial display in Figure 5, then in order to view the data in a descriptive/numerical format, the bar chart must be selectively chosen (which then hides the map). Visually combining spatial and numerical data (in the form of charts) is important for numerical-spatial problem solving. The importance of this convergence by looking as far back as John Snow's methodology for discovering the source of the cholera outbreak in London during the 19<sup>th</sup> century [52], was previously described. Using both spatial and numerical display to map death counts to spatial objects, Snow was able to quickly determine the source; a contaminated water pump near a popular brewery.

Richard developed an OLAP-GIS system for web-based access to End-Stage Renal Disease (ESRD) [53]. The system, known as SIGNe, is an add-on to the Renal Epidemiology and Information Network (REIN) in France. The Multi-Source Information System (MSIS) is the original component of the REIN. Here individual patient information related to End-Stage Renal Disease could be entered, processed, and loaded into a database for later review. The MSIS is currently running into 8 of 22 French regions. The architecture is n-tier, with universal clients representing the first tier, a web-server and business logic features representing the second tier, and the databases representing the third tier. SIGNe was built to integrate with MSIS-REIN. The purpose was to give clinicians a tool for health care decision making in relation to ESRD.

The backbone of SIGNe is the OLAP data warehouse. Here clinical (patient) information is received from an MSIS database. Feeds to the data warehouse are done twice a year. The other data loaded into the data warehouse is external data (outside of the MSIS architecture). This information mainly contains census and spatial information for SIGNe. The spatial and numerical information are then combined together in the data warehouse. This decision to

combine the numerical and spatial information into the data warehouse is the same as Bedard's system. Once combined in the OLAP data warehouse, the data is accessed through the SIGNe querying engine. The data warehouse is also linked to a data mining and modeling component. This resembles the SOVAT architecture where the Integration Engine connects to the modeling subsystem and enables dynamic user modeling and data mining features (this will be discussed in more detail in section 3.2.1).

SIGNe contains a web-server that enables for online access. The interface of SIGNe can be accessed through a universal client. The user is able to specify dimensions and attributes on which to query. Data is displayed in either a chart or a graph form. The interface does not appear to contain true OLAP-GIS functions such as drill-out. In addition, the interface does not appear to support direct manipulation of spatial objects. Rather, traditional OLAP front-end features such as tabs and menu selections are used to drive the query formulation process.

The final OLAP-GIS system to be discussed is developed by Bapna [54] for analyzing motor vehicle accidents. Like SIGNe, the system is also web-based. The system is called Maryland Spatial Analysis of Crashes (MSAC). The MSAC architecture contains 5-tiers: presentation layer, web-server layer, map server layer, GIS (gateway) layer, and database layer. The OLAP cube is located in the fifth tier and retrieves data based on requests from the GIS gateway in the fourth tier. The presentation level interface supports mapping, drilling, reporting, and even email functions. With the mapping functions, the user is able to view the specific crash locations. The drilling functions supported by OLAP then enable the user to examine data by different dimensions such as county, and streets/routes. The OLAP functions are specified by the user drop-down lists and check box windows. Unlike SOVAT, there is no ability for direct manipulation on the map as queries are only specified by these lists and check boxes. In

addition, MSAC does not have any functions such as “drill-out” that demonstrate a true synergy between OLAP and GIS capabilities.

An example of a non OLAP-GIS system for numerical-spatial problem solving is the Epidemiologic Query and Mapping System also known as EpiQMS. EpiQMS is an example of a system that combines relational databases with GIS [55]. The system was developed by the University of Washington and Washington Department of Health in an effort to increase the availability of up-to-date online health-related data to three different types of users: the general public, public health professionals, and clinicians. One of the main focuses of the project was 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 5 [55]. 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 [55]. 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. 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. Recently, EpiQMS has been featured by the Pennsylvania Department of Health and made available through their website<sup>9</sup>. The department hopes to develop this tool and make it a valuable online health information tool for the general public. The Pennsylvania DOH has expanded the system to include more dynamic mapping capabilities and more statistical measures. The system still has a traditional database format, while the interaction with the mapping is extremely narrow, due to limitations with navigation caused by the back-end architecture of the system. This results in the ineffectiveness to create even numerical routines. Thus this system is very weak in regards to addressing numerical-spatial problem solving.

Hernandez et al [56] have developed the commonGIS system for environmental decision support in Nicaragua. The system is actually a GIS DSS generator that provides spatial analysis, display, and interactivity through an elaborate graphical user interface. The system has been coupled with the ‘SPIN!’ system [57] 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. The major drawback of the ‘SPIN!’ system and the commonGIS generator is that they both rely on a relational database backend which most likely impedes analysis. There are plans to add OLAP to commonGIS in the near future.

The Nehme project for land evaluation [58] is another system that uses the GIS component to simply display information rather than incorporating it into the process of query development.

---

<sup>9</sup> <http://app2.health.state.pa.us/epiqms/> (11/01/04)

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 [12, 13, 15] , 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) [59]. 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. The system combines GIS presentation with relational database architecture. While there are plans to add interactive data mining components, the system is more of an automated monitoring system with spatial display, rather than a decision support system that supports interactive query development.

Certain government agencies have also created spatially-related decision support systems. The Health Resources and Services Administration (HRSA) has developed a geospatial data warehouse for use by the general public [60]. 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)<sup>10</sup> software and a relational database to allow users to log on and create maps concerning reproductive health issues, such as infant mortality, birth weight, and fertility [25]. 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 [61, 62]. 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<sup>11</sup>, which is part of their desktop GIS suite. The system uses relational database architecture rather than the multidimensional structure of OLAP. It is more of a stand-alone GIS system than a specific DSS, and thus does not support problem specific modeling or any intelligent AI features. It does however use data mining techniques such as numerical clustering (by incorporating SaTScan software, which was developed by Kulldorff [63]), in an effort to find interesting patterns within the data.

Table 2-3 summaries the four OLAP-GIS decision support systems: MSAC, SIGNe, ICEM-SE, and SOVAT. The factor that separates SOVAT from the rest is its interface features. As will be shown in the next chapter, SOVAT is the only system that contains 'true' OLAP-GIS functions such as drill-out. In addition, it is the only system to enable for direct manipulation of spatial objects such as counties or municipalities. For example, in SOVAT, the user can select a county by clicking on the map and then drill-down directly from that county to view its

---

<sup>10</sup> <http://www.esri.com/software/arcgis/arcims/index.html> (11/01/04)

<sup>11</sup> <http://www.esri.com/software/arcgis/arcview/index.html> (11/01/04)

municipalities. The other systems require menu or check box selection outside of the spatial area on the interface. In addition, the fact that SOVAT is the only system in which the spatial data is not combined in the OLAP cube does not interfere with the ability to provide OLAP-GIS capabilities. In SOVAT, the spatial and numerical data are combined in the integration engine and displayed on the user interface.

**Table 2-3: Comparison of the OLAP-GIS DSSs.**

	<b>Environment</b>	<b>Data in OLAP Cube</b>	<b>Modeling/Data Mining Capabilities</b>	<b>Manipulation of Spatial Objects through Interface</b>	<b>OLAP-GIS Functions</b>
MSAC	Web-based	Combined – spatial and numerical	None	None – querying through lists and checkboxes	None
SIGNe	Web-based	Combined – spatial and numerical	Yes (specifics not mentioned)	None – querying through lists	None
ICEM-SE	Web-enabled (with plug-in)	Combined – spatial and numerical	Yes (specifics not mentioned)	Can select map, but Drill-down, drill-up, and slice is done outside of map area	None
SOVAT	Desktop	Separate – numerical only	Community creation, Spatial clustering	Drill-down, drill-up, drill-out. Spatial slice	Drill-out

### **3. COMBINING OLAP AND GIS FOR DECISION SUPPORT**

#### **3.1. Introduction and Background**

A system that combines Online Analytical Processing with GIS will create a potentially powerful architecture for numerical-spatial problem solving. By combining these two technologies, a system inherently has access to their characteristics and functions, but also the ability to possess unique functions that are only possible through their merger. As explained, few OLAP-GIS systems exist. Other projects have focused solely on using GIS with spatial and statistical analysis and do not attempt to the change from the two-dimensional data subsystem component that is compatible with these geospatial DSS generators. In addition, individuals who claim they are developing knowledge discovery decision support systems generally incorporate data mining as a stand-alone software application rather than as a coupled DSS generator for the decision support process. Thus an integration of even just two of the three components (OLAP, GIS, and data mining) is rare in the DSS literature. OLAP contains many unique functions that allow for efficient and powerful data exploration. Drill-up, drill-down, slice, and dice can be invoked by the OLAP engine, which allow for the data to be accessed from the multidimensional cube, and then quickly display the results in either tabular or graphical form. By combining OLAP and GIS<sup>12</sup>, there is the ability to incorporate a new OLAP function, “drill-out” that can be applied to geospatial objects and subsequently enhance the decision support process. Figure 3-1 and Figure 3-2 show this process. A spatial object (county in this case) is selected. The “drill-out” button is selected which identifies and then includes the surrounding spatial objects within the analysis. The GIS component allows for the bordering spatial objects to be found (through

---

<sup>12</sup> And enhancing these two DSS generators with the appropriate DSS tools such as a visual programming language and necessary third-party components

coordinate calculation) and highlights the selected items, while the OLAP function performs the aggregation and displays the results.

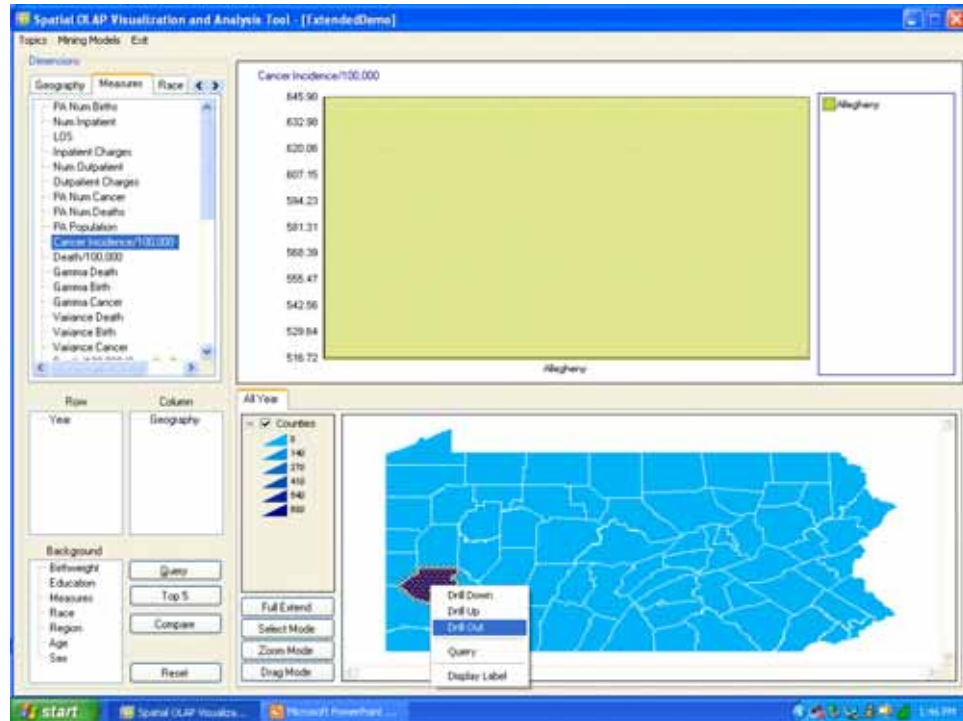


Figure 3-1: "Drill-Out" function in an OLAP-GIS system.

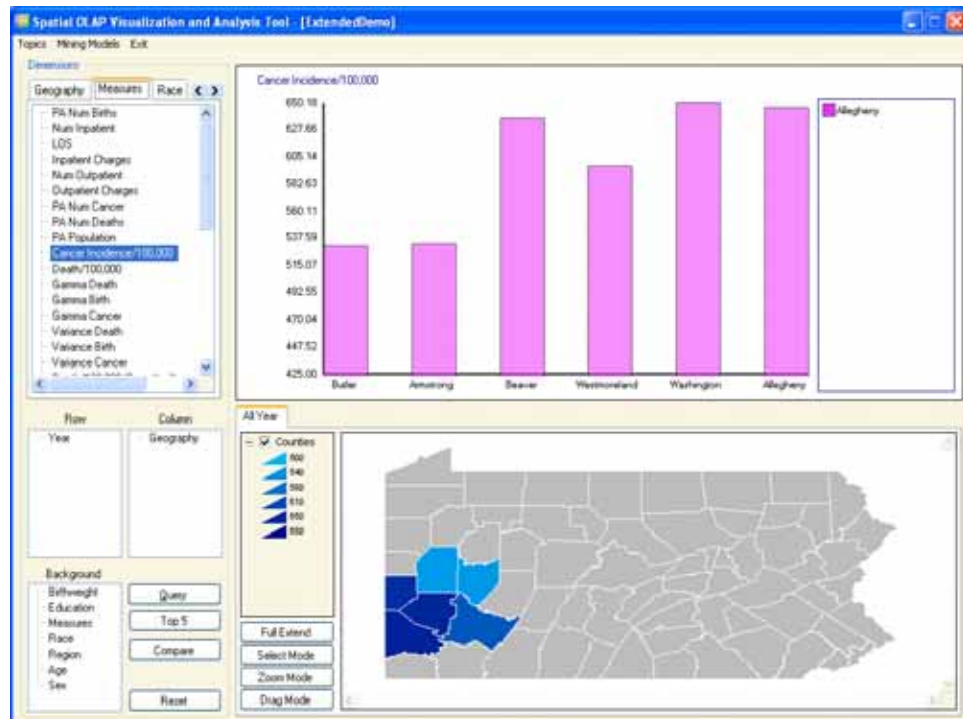


Figure 3-2: "Drill-Out" for boundary detection.

This synergy between OLAP and GIS provides for a form of spatial aggregation that can be very useful for decision support purposes such as community health assessment analysis. For example a researcher interested in examining the rate of incidence for a particular disease can perform numerical-spatial query development by using OLAP and GIS. For example, he/she can view one county, and then by performing the “drill-out” operation, instantaneously see the affects of the disease on the county and its neighbors. This will allow the expert to assess if the disease is a broad issue or a local one. Without being able to drill-out, the user would need to perform this spatial aggregation by him/herself. This would involve browsing through the geography dimension, identifying the relevant spatial objects, highlighting them, and then sending the query to the OLAP engine. This in essence is like performing drill-out by using only OLAP, which works but it not as easy as with adding GIS to this operation. Drill-out has been demonstrated by many OLAP vendors, but the function is not inherent to the technology, rather the process of “drilling out” is performed by the user through the manual selection of relevant attributes [64]. Hence there is no recognized OLAP function for drilling out on a dimension. More importantly, there have been no examples of using OLAP and GIS for creating a drill-out function. Thus the drill-outs demonstrated in the literature have been purely numerical without any added spatial component (such as the recognition neighboring spatial objects). The enhancement of the drill-out feature into an OLAP-GIS DSS enhances the development of numerical-spatial queries for the decision support process and gives the user a sense of synergy between OLAP and GIS.

The ability to drill-out with an OLAP-GIS DSS challenges the assertion from Gao et al [49] that it is difficult to create numerical-spatial routines within one step. This can be seen by enhancing the drill-out operation to include dimension-specific attributes. For example, two drill-

out buttons can be added to the interface; one that drills-out on the male attribute of the *Sex* dimension, and one that drills out on the female attribute. This can be seen by examining the previous drill-out example. As mentioned, drill-out starts by the system identifying (through a coordinate calculation process) the spatial objects that border the currently selected object. This here, is comparable to the example in Gao et al of using buffering for spatial query development (see page 7 of [49]). If the user selected the ‘Males’ drill-out function, the OLAP (or numerical query) process occurs by spatially aggregating the relevant objects and performing the calculation by slicing on the male attribute. The same process can occur by pressing the female button. While the additional drill-out buttons for each numerical query might not be appropriate from a usability standpoint, the proof of concept is shown that creating a drill-out function by combining OLAP-GIS does allow for spatial and numerical components to be combined within a single process. Correspondingly, “drill-in” could also be added which could reverse the aggregation of the spatial objects. These two functions at least exemplify the synergy between OLAP and GIS for the user during the decision support process.

As mentioned previously, the combination of OLAP and GIS into a DSS inherits the capabilities that these two technologies provide independently (as well as the additional drill-out/in functionality). Table 3-1<sup>13</sup> builds upon Table 2-1 by adding an OLAP-GIS system. Combining these two technologies fills in many of the functionality gaps (listed on the top) from the individual generators.

---

<sup>13</sup> Reprinted from International Journal of Medical Informatics, In Press, Scotch M Parmanto B, Development of SOVAT: A numerical-spatial decision support system for community health assessment research, 2005, with permission from Elsevier.

**Table 3-1: Traditional DSS generators versus an OLAP-GIS system.**

	Interesting Patterns/ Knowledge Discovery	Large, Complex Data Sets	Multidimensional View/ Navigation	Statistical Analysis	Spatial Presentation	Visual Charts	Spatial Analysis	Numerical- Spatial Problem Solving
Statistical Software				X		X		
Data Mining	X			X		X		
GIS Software					X		X	
Relational DB		X						
OLAP		X	X			X		
<b>OLAP-GIS</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>	<b>X</b>

OLAP addresses the need to handle large and complex data sets and allow for comprehensive data exploration through its multidimensional cube model. It also supports many off-the-shelf desktop clients for visual display. GIS can perform statistical analysis, as well as perform spatial presentation and spatial analysis. The other characteristics, Interesting Patterns, Statistical Analysis) need some further explanation. It has already been mentioned that OLAP is only designed for simple mathematical functions, such as count, sum, and mean. Thus strong statistical functions that are provided by statistical packages such as SAS or SPSS are not inherently provided. A few examples that tackle this problem have been addressed, such as coupling statistical DSS generators within the problem processing subsystem of the decision support architecture. Other solutions are to pre-calculate the data offline and then process the OLAP cube. Another solution to this problem can be addressed by utilizing DSS tools such as visual programming languages to access their mathematical libraries and calculate these numbers on-the-fly. A final solution is to create calculated measures either before the processing of the OLAP cube or by the user during run-time that can be inserted into an ad-hoc query. Either way, the user has the choice of using or omitting the statistical functions during system utilization. It is important to note that that calculated measures are provided by the querying (access) language of the OLAP DSS generator. For example MultiDimensional Expressions (MDX) is the

language used by Microsoft SQL Server 2000 to perform operations on the OLAP cube [65]. It is through this language that the OLAP function (drill-up, drill-down, etc) as well as the creation of calculated measures can be performed. Calculated measures can be used for incorporation of important decision support statistical functions such as spatial smoothing and age-adjustment.

The final characteristic that needs to be addressed is the ability to use data mining and knowledge discovery in order to find interesting patterns in the data. OLAP and GIS as independent DSS generators do not support data mining functions. It was mentioned before that there have been a few popular solutions to this problem. The easiest is to take the data from the data warehouse, preprocess it and run it through a stand-alone data warehouse application such as SPSS Clementine. Another method is to make a statistical software application a DSS generator (coupled) within the DSS architecture. This typically involves integrating it with the other generators within the problem processing subsystem component of the DSS. A final scenario is to utilize the capabilities of some OLAP DSS generators that allow for the creation of data mining models. SQL Server 2000 allows for clustering and decision tree models to be created against an OLAP cube. The new 2005 version<sup>14</sup> has added more mining capabilities. The models can then be utilized by the user through the development of DSS tools such as a visual programming language.

The characteristics listed in Table 2-1 are a belief of what features are important to a numerical-spatial decision support system. They are based on the components discussed by Gorla in measuring perceived ease of use and perceived usefulness of an OLAP system [66]. The true acceptance of an OLAP-GIS system for decision support will be realized through objective evaluation of the features listed in Table 2-1 versus currently utilized technology. It is

---

<sup>14</sup> <http://www.microsoft.com/sql/2005/productinfo/top30features.asp> (11/01/04)

anticipated that the unique combination of OLAP and GIS will provide users with a powerful for numerical-spatial decision support.

### **3.2. Development of SOVAT**

The Spatial OLAP Visualization and Analysis Tool (SOVAT) system was developed for numerical-spatial decision support. The system combines OLAP and GIS technology to create a unique and powerful tool for answering numerical-spatial problems. For the purposes of this research, a community health assessment focus (for Pennsylvania) will be used as the case study. It should be noted that the system is capable of handling any domain of interest, with only the data subsystem component needing to be altered. The architecture of the system can be seen in Figure 1-6.

Data necessary for performing community health assessments was collected from three main sources: The Pennsylvania Department of Health<sup>15</sup>, The Pennsylvania Healthcare Cost Containment Council<sup>16</sup>, and The United States Census Bureau<sup>17</sup>. 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*. An illustration of some of these dimensions can be seen by looking at the multidimensional cube in Figure 1-5. A few of these dimensions need further explanation.

---

<sup>15</sup> <http://www.health.state.pa.us> (11/11/04)

<sup>16</sup> <http://www.phc4.org> (11/11/04)

<sup>17</sup> <http://www.census.gov> (11/11/04)

The diagnosis dimension is based off of the World Health Organization's International Classification of Disease 9-CM coding hierarchy<sup>18</sup>. 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; state-wide, county level, and municipality level. The hierarchy goes from state-wide 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. In order to ensure confidentiality, the hierarchy was limited to the municipality level (the census bureau actually provides data down to the block level). While Pennsylvania was used to model the geography dimension, it should be noted that the flexibility of the system allows for any type of geography to be used. For example, given nation-wide 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 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, the Race dimension was created 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 analysis. For further information, please see [67].

---

<sup>18</sup> <http://www.cdc.gov/nchs/about/otheract/icd9/abtcd9.htm> (11/11/04)

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’<sup>19</sup> community health OLAP cube. The tables were housed in the data subsystem component of SOVAT.

### **3.2.1. The Integration Engine**

The combination of OLAP and GIS occurred through the development of the integration engine (created using the .NET platform). DSS tools are not only used to construct DSS generators, but also to provide an interface between them. The integration engine enables both the OLAP and spatial functions to act on both the numerical and the spatial data. This part of the SOVAT architecture is the most vital because it couples the two technologies and enables for both spatial and numerical tools to be synergized. This is the area where many developers struggle; integrating the DSS generators and tools together in order to create an effective system. 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. Thus it is essential to create and utilize DSS tools that interface both components and are able to integrate them effectively. The platform 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 the integration engine.

### **3.2.2. OLAP and GIS 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-

---

<sup>19</sup> OLAP cubes can be combined in Microsoft SQL Server to create what are called ‘Virtual’ Cubes. The cube is a snapshot of the individual data marts and not a fully processed stand-alone cube. To the user, however, it represents the combined data form all the sources.

use yet powerful interface. Two important DSS tools were added to allow for numerical-spatial problem solving. These consisted of numerical and spatial 3<sup>rd</sup> party component software. 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 is used to display the results from the OLAP cube in the form of a bar chart. A detailed description of SOVAT's user interface will be discussed in the next chapter.

### **3.3. Features in SOVAT for Numerical-Spatial Problem Solving**

The purpose of this research is to construct an OLAP-GIS decision support system that allows for numerical-spatial problem solving within a single interface. In chapter 2, it was outlined which technologies are commonly used in a data-oriented decision support system. After analyzing each of these independently, the important features for numerical-spatial problem solving were identified. A culmination of this information can be found in Table 2-1. The assertion is that combining OLAP and GIS in a decision support system will provide the necessary features needed for solving numerical-spatial problems. This will be realized through the creation of numerical-spatial routines within a single system. The unique OLAP-GIS system, SOVAT, contains these necessary features: knowledge discovery, handling of large and complex data sets, multidimensional data exploration, statistical analysis, spatial presentation, visual presentation, and spatial analysis. The culmination of these features leads to the ability to perform useful numerical-spatial routines. Each of these features as applied to SOVAT will now be addressed.

### **3.3.1. Knowledge Discovery**

As discussed in chapter 2, the ability to discover interesting patterns within the data is becoming an integral focus within decision support system research. Classification techniques that tie in both numerical and spatial information can prove very valuable during the decision support process. As mentioned, when integrating OLAP into an On-Line Analytical Mining (OLAM) environment, there are certain functions that are appropriate for this data model and ones that are not. The appropriate data mining techniques (clustering, decision rules) have been integrated within SOVAT. Data mining models have been developed that allow for numerical clustering of spatial objects. For example, with this feature, a user can answer “*What counties are similar in relation to male childhood cancer hospitalization?*” This could then lead to further data exploration by examining specific counties in a “high” cluster. This data mining process combines both numerical and spatial elements. For example, the desire to create clusters based on various statistical attributes is purely a numerical process. However, the need to then color code the corresponding geographical (county) elements on the map to highlight their distance to one another is a spatial process.

Currently, the user can select a pre-developed data mining model from a menu at the top of the screen. Further development intended for expert use will aim at allowing users to construct their own mining models for clustering and association rules generation.

### **3.3.2. The Handling of Large, Complex Data Sets**

The ability to handle large and complex information is one of the inherent characteristics of OLAP. Any system that contains OLAP within the data subsystem component will be able to do this. OLAP provides three types of models for data storage: Relational OLAP (ROLAP), Multidimensional OLAP (MOLAP), and Hybrid OLAP (HOLAP). The differences between

these techniques are beyond the scope of this analysis and thus will not be addressed (please see [67] for further explanation). For the most part, regardless of which technique is chosen, the OLAP environment will allow for massive amounts of data to be stored and easily processed for detailed exploration.

### **3.3.3. Multidimensional Data Exploration**

The ability to explore multidimensional data (conceptually in the form of a data cube) is *the* inherit component within OLAP. Once the star schema is developed, the multidimensional cube is processed and created, allowing for ad-hoc exploration through unique OLAP features: drill-up, drill-down, slice, dice, and pivot. The SOVAT system contains the ability to provide all of these OLAP-specific functions and allows for rich ad-hoc data exploration. This process is driven by the OLAP engine provided through SQL Server's Analytical Services. As a stand-alone application, OLAP can be used for powerful and quick numerical query development through interactive drilling and slicing of the multidimensional cube.

### **3.3.4. Statistical Analysis**

OLAP has always been criticized for only supporting simple mathematical functions, such as count, sum, min, and max. Users have seen its ability to handle large amounts of data and allow for quick and powerful data exploration as outweighing this negative characteristic. As was highlighted in chapter 2, statistical analysis is an important feature in solving numerical-spatial problems. Thus it is an important feature that must be included in the decision support system. One manner in which to do this is to couple the problem processing or data subsystem components with a statistical DSS generator such as SAS or SPSS. This however, can reduce processing time and impede problem solving. A different solution was taken with the development of statistical analysis in SOVAT. Important statistical functions were implemented

(such as spatial smoothing and age adjustment) as calculated measures within the OLAP cube. The measures were developed using the Multidimensional Expression Language (MDX) which is used by the SQL Server OLAP engine for access to the OLAP cube. Based on the multidimensional environment, the calculated measures became attributes of the *Measures* dimension (not described earlier since it is produced inherently by SQL Server and not the developer). The user can then insert this measure into any query in order to use these special statistical calculations. For experienced users, SOVAT will soon enable for the ad-hoc development of statistical measures. At the moment, all statistical measures are pre-defined and located in the Measures dimension. For the purposes of community health analysis, statistical smoothing and age-adjustment were the primary measures added. Additional statistical functions can be added to SOVAT when needed.

### **3.3.5. Spatial and Numerical Presentation**

Spatial problems can not be addressed without spatial presentation. Geospatial Information Systems (GIS) provide users with the ability to integrate vector or raster layers (roads, lakes, houses) on top of one in another in order to assemble a digital map. To the GIS user, the spatial display *is* the system. Integrating GIS within SOVAT allows spatial problems to be analyzed. The spatial software which was embedded into the interface enables for spatial display of digital maps. Conversely, the numerical (presentation layer) software enables the data from the OLAP cube to be displayed. Traditionally this is done in the form of a bar chart; however the user has the ability to switch to other charting methods such as a line graph or pie chart. There is potential for a decomposition tree (data presented in a hierarchical format) to be added as well. The ability to display data in a numerical and spatial format is important, but there needs to be synergy between the two components in order for effective query development to occur. For this

purpose, SOVAT's integration engine couples the spatial and numerical information and allows the resulting query to display in both the bar chart and the digital map. When a query is performed interactively on one visual component, the result is updated on both. This is a different approach for spatial and numerical query development from the one taken by Gao et al [49]. Their technique is to completely separate the two processes and present the information in completely different windows. Thus with their system, a numerical query has no impact on the spatial display. It is anticipated that a more integrated approach, like SOVAT, will enhance numerical-spatial problem solving.

### **3.3.6. Spatial Analysis**

The ability to present spatial information via spatial display is only part of spatial problem solving. Effective spatial analytic techniques need to be available to the user in order to construct queries. Gao et al [49] demonstrates some useful functions for this process. In chapter 2, some of these functions were highlighted, including buffering and network analysis. The spatial software in SOVAT provides for spatial analytic functions to be available. The spatial functions used during community health analysis vary depending on the problem. For example, after identifying areas of high incidence of a particular cancer, the expert might have suspicions that the increase is caused by an environmental exposure. He/she may then want to know the distance of this population from a popular river. Spatial techniques such as network analysis and buffering would be useful for this research question. Color gradation of numerical values is also an important approach needed for spatial analysis. As mentioned, this is different than numerical clustering of spatial objects, which is a data mining approach. However, using color grades to identify numerical trends in spatial objects is a useful research tool (i.e. "Which counties are in

the same color region for a particular numerical measure?”). SOVAT contains this ability, and the results can be simultaneously displayed on the numerical graphs.

### **3.3.7. Numerical-Spatial Routines**

As frequently mentioned, the research focus is on combining OLAP and GIS for numerical-spatial problem-solving. It is the belief that this unique integration will enhance this process. This work builds upon the notion of combining numerical and spatial components introduced by Gao [49]. They provide a framework for this process by introducing different layers within a DSS architecture. Gao focuses on identifying specific DSS generators for this process, rather than defining a blueprint for how this system is to be constructed. After all, it is the DSS generators that provide the technology for the decision support process. OLAP and GIS are the fundamental DSS generators for numerical-spatial problem solving. Previously outlined were the important features that they provide as necessary for the decision support process. The combination of all these features is necessary for creating a powerful environment for numerical-spatial problem solving. Additional functions have been realized after the integration of OLAP and GIS. The drill-out/drill-in function was highlighted, which combines spatial and numerical techniques. The benefit and power of combining OLAP and GIS for decision support was previously shown. The next three chapters (4, 5, and 6) will focus on the studies related to this work.

## **4. USABILITY ASSESSMENT OF SOVAT**

### **4.1. Introduction**

This chapter details the first study involving the Spatial OLAP Visualization and Analysis Tool (SOVAT) - a usability analysis. The uniqueness and novelty of the system presents the potential for many issues during human-computer interaction (HCI). Thus, before this system can be evaluated in a summative study, it must be first deemed 'usable'.

### **4.2. Background**

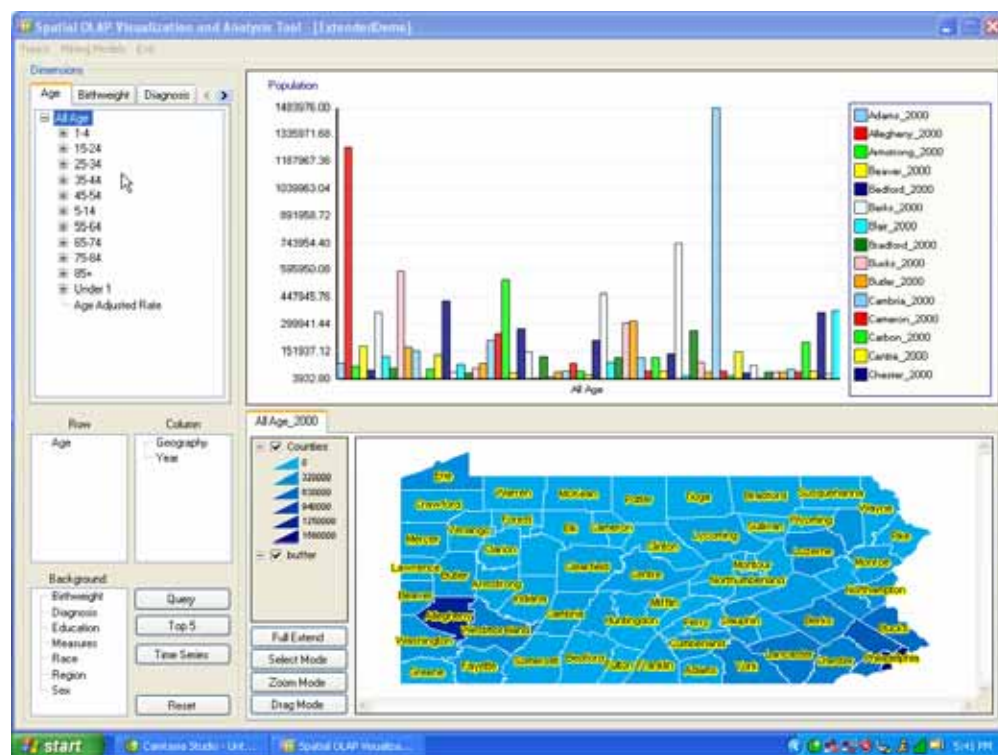
Since there are few realized OLAP-GIS systems, a usability study examining the issues that this synergy might have in relation to HCI, has yet to be explored. GIS has generally been perceived as a complex technology, since most users prefer to only use it for visual display of spatial information. Many of the functions that distinguish GIS from a simple viewer that displays spatial information (such as direct map manipulation, buffering, geocoding, etc) are never utilized. OLAP is considered even more complex than GIS. Despite being able to support certain eloquent front-end clients (such as ProClarity) the notion of a multidimensional cube with dimensions, attributes, and special drilling methods is much more daunting from a conceptual standpoint than traditional flat-file relational tables. The difficult task with this study is to take a very powerful and complex system and make it into a very powerful user-friendly system. This goal can not be met without a rigorous usability study.

The motivation behind SOVAT is to create a decision support system that is usable for any community health researcher regardless of computer experience. The challenge is to utilize the

powerful features of OLAP and GIS while creating an interface that is not only easy to use but can also address complex community health research questions.

### 4.3. SOVAT Interface

Figure 4-1 shows the original version of the SOVAT interface (at startup mode). This graphical user interface (GUI) supports drag and drop as well as other direct manipulation actions.



**Figure 4-1: Original SOVAT interface at ‘startup’.**

The interface can be divided into four regions: the dimension tabs (top left), charting area (top right), row and column lists as well as special functions (bottom left), and finally the map area (bottom right). The dimension tabs represent the types of information stored in SOVAT. The dimensions in SOVAT represent the combination of health and population data sets, and are: *Age, Birth weight, Diagnoses, Education, Geography, Measures, Race, Region, Sex, and Year.*

Each dimension has its own tab. Dimension tabs can be accessed by using the arrow button by the tab names to scroll through the different tabs. Attributes or elements within each dimension are represented in a hierarchical tree-like format. Attributes can be expanded or collapsed depending if they have a '+' or '-'. The hierarchical layout of the dimensions reflects the concept that the data is stored within a multidimensional OLAP cube. The use of tabs for representing dimension hierarchies is a traditional OLAP front-end feature. The use of a chart to display descriptive data is another popular visualization technique in OLAP. The final OLAP characteristic are the row, column, and background list boxes. The row and column list boxes allow the user to define which dimensions should be included in the query and how the results should be displayed on the bar chart. Users can drag a dimension name from the background list into either the row or column list before a query is submitted. The map area represents the traditional GIS interface features and contains technology for addressing spatial problems. It will display the same (numerical) results as the bar chart with the data being displayed via color gradation and defined on the map's legend.

There are many ways to submit queries using SOVAT. The first method is to select the necessary attributes from the dimension tabs and drag and drop them onto the charting area. Multiple attributes from multiple dimensions may be selected at once by holding down the shift or control key. A similar method of querying involves selecting the appropriate attributes, but instead of dragging and dropping, the user simply selects the query button by the background list. Querying can also be done through direct manipulation of either the chart or the map. To query with the chart, the user needs to select a bar and right click. A menu will appear that displays the common OLAP functions 'drill-up' and 'drill-down'. This will allow the user to perform these functions on this specific data cell. For example, the second highest bar in Figure

4-1 is Allegheny County (second to left). Selecting this bar and choosing drill-down will produce a query that displays the municipalities (which are the hierarchical children) of Allegheny County. Querying with the map can be done in a similar fashion. Selecting a map element, such as a county, and then double clicking on it, automatically performs a drill-down function on that county. In addition, right clicking after selecting a map element displays additional options. In addition to drill-down and drill-up, there is also ‘drill-out’, which is a custom OLAP-like function included in SOVAT that is not present in standard OLAP technologies. Rather than moving to a higher or lower level of geographic granularity, drill-out expands outward and performs aggregation on bordering counties or municipalities. This function is essential for community health assessment experts who need to analyze a specific community in relation to its bordering areas. The last option with the map is the ability to define a custom community. For the researcher, a community might be more than a specific county or municipality. Thus, he/she needs to analyze a group of spatial areas as a ‘set’. Being able to define this area and use it as a single entity within the interface will potentially facilitate numerical-spatial problem solving. For example, if the researcher needs to analyze northwest Pennsylvania as Crawford, Warren, and Erie counties, he/she can select these counties, hit right click, select “Save as Community” and type a name for the area. This will then aggregate these counties into one unit and allow the researcher to analyze this newly defined community in this manner.

The last method of querying with SOVAT is to use the custom “Top 5” function. Depending on which dimension is displayed in the column list, hitting the Top 5 button will produce a rank listing of the dimension’s attributes. For example, if Geography is in the column list and Erie county is highlighted in the Geography dimension tab, the top 5 municipalities of Erie county (in relation to whichever health measure is being analyzed) will be displayed. This

is an important function for community health assessment researchers because it allows them to quickly get a sense of the main health issues for a given community, or the main geographical areas that are most significantly affected by a specific diagnosis and/or health measure.

#### **4.4. Methodology**

A think-aloud assessment was implemented to identify system usability issues. SOVAT is intended for professionals who analyze health and population data for the purpose of learning about health factors within a defined population. For this study, both future users and representative users were recruited. Future users were defined as students within the University of Pittsburgh's Graduate School of Public Health. Within this school are such departments as: Behavioral and Community Health Sciences, Biostatistics, and Epidemiology. Representative users are working professionals who routinely work with health and population data to perform community health analysis. This might include individuals within local/state health departments such as biostatisticians and epidemiology data managers; data analysts within business or academic research settings; university faculty/researchers; data analyst in non-profit foundations; and data analyst in healthcare institutions (hospitals, HMOs, etc.). Ideally only representative users would be involved in the study, however due to the scarcity of these individuals locally as well as the time commitment involved in the study session (which means time away from their workplace), it was determined that a mix of professionals and students was more practical.

##### **4.4.1. Recruitment and Setting**

Nine graduate students and six professionals participated in the study which took place in the conference room at the Center for Biomedical Informatics, on the campus of the University of Pittsburgh. Five rounds of testing took place. The graduate students were recruited via fliers

that were posted around their school (**Appendix A**). The professionals were identified by contacts and were sent a formal recruitment letter (**Appendix B**). Separate study times were scheduled for each of the participants. In the conference room was a laptop with an external mouse and external microphone. The conference room also had a pull-down projector screen for instructional purposes.

#### **4.4.2. Study Procedures**

Before entering the conference room, participants were asked to complete the informed consent form (**Appendix C**) as well as a short background questionnaire on their computer experience (**Appendix D**). The usability study lasted approximately an hour and a half. Once in the conference room, the participants were shown a pre-recorded 8-minute instructional video that served as the introductory script for using the system. The content of the video, including the facets of the interface and the methodologies for producing queries, was deemed appropriate for use in a usability study by one of the co-investigators (VM) who is an expert in Human-Computer Interaction (HCI). After watching the video, the participants were not allowed to ask any additional questions related to using SOVAT. They were then instructed on the procedure of ‘thinking aloud’ and asked to do this as they interacted with the system.

The participants were given five problem solving tasks to answer using SOVAT (**Appendix E**). The tasks represented realistic community health assessment problems, and were deemed appropriate by a co-investigator (RKS) who is an active community health researcher. They consisted of performing local and state-wide comparison of geographic areas, ranking of diseases or geographic areas based on health measures, and defining and comparison of customized geographic communities. The tasks were not randomized, since it was preferred to have the perceived easier tasks first. Camtasia screen capture software was used to record their

interaction while the external microphone captured their verbal thoughts. Once the participants completed the 5 tasks, they were asked to complete two usability questionnaires.

#### **4.4.3. Objective Measurements**

In order to identify usability issues, the following four variables were considered:

- Time to complete each task – This measure was defined by the time between when a participant finished reading the question to when the participant indicated he/she was done. The use of screen capture software allows one to measure the participant's time for each task. The desired benchmark for this study was less than 5 minutes to complete a task. This screen capture method is also non-intrusive.
- Erroneous Action – An erroneous action was defined as an action that did not get the user closer to their goal of solving the problem. The desired benchmark for this study was less than 5 erroneous actions per task.
- Problem Space – A problem space was defined as an action that represents a different method of solving a task sub-goal than what was previously tried. This symbolizes a back track in the user's problem space. The desired benchmark for this study was less than 3 problem space approaches per task.
- Answer to Problem – An answer was defined as the action of the participant verbalizing an answer to all the questions in the task followed by saying that they were 'done'. The answer did not have to be the same as what was currently being shown on the screen at the time. The participant had to answer all parts of the question correctly in order to correctly answer the task. The desired benchmark for this study was a correct answer per task.

Thresholds were determined prior to the study based upon an expert's completion of the tasks. Based upon the measurement scores of the expert, the thresholds were established by estimating how a beginner might perform.

#### **4.4.4. Subjective Measurements**

Usability issues were also identified subjectively. A post-study questionnaire called the IBM Post-Study System Usability Questionnaire (PSSUQ) (**Appendix F**) was used. This is mainly a close-ended questionnaire that has been found to be both a reliable and valid instrument for lab-oriented usability evaluation [68]. The PSSUQ utilizes a 7-point Likert Scale format with lower numbers indicating higher levels of satisfaction. The questionnaire is designed to analyze across three categorical areas: system usefulness, information quality, and interface quality. System usefulness corresponds to the user's belief in the system to improve job their performance. It is considered one of the most important psychometric variables because it has been closely linked to user acceptance and adoption of information technology [69]. Information quality corresponds to the user's belief in the system to provide them with helpful and important information to complete the tasks. This could include help screen, help messages, and clear display of information content. The final category is interface quality which relates to the evaluation of the user interface layout.

Besides the PSSUQ, an additional more open-ended questionnaire was also used to record opinions about the best and worst aspects of the system (**Appendix G**).

#### **4.5. Results**

The usability study consisted of five rounds, with three participants per round (for all total of 15 participants). The participants included students from the University of Pittsburgh's Graduate

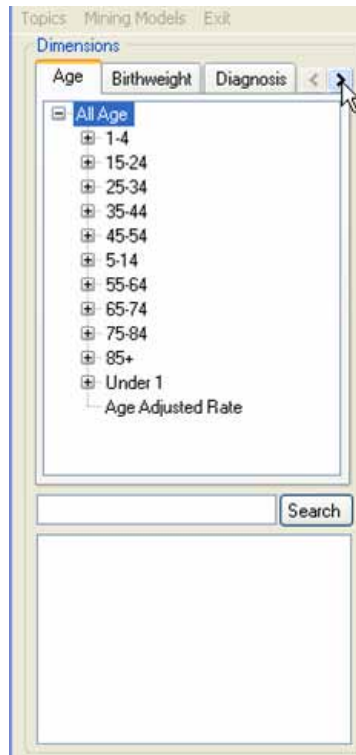
School of Public Health and community health professionals from either local health departments or the University of Pittsburgh. The students participated in the first two rounds and the final round (round 5) of the usability assessment. The professionals participated in the third and fourth rounds of the study (it was preferred to have professionals for the final round, however this did not occur due to difficulty in recruitment). After each round, the investigators reviewed the objective and subjective results and made changes to the interface based on the findings.

#### **4.5.1. Interface Changes Made**

As described in the last chapter, the interface can be divided into 4 components: dimension tabs, chart, map, and row/columns and special functions. These components will be described separately for the purpose of detailing the changes made to the interface.

##### **4.5.1.1. Dimension Tabs**

In round 1, some participants encountered difficulty in finding a diagnosis attribute within the diagnosis hierarchy. The diagnosis hierarchy contains branches up to three levels deep (e.g. Respiratory System → Chronic Obstructive Pulmonary Disease → Asthma). One participant listed the fact that there was “No ‘Search’ option” as one of the worst things about the system. In order to enhance the usability of the dimension tab, a search option was added under the dimension tab (Figure 4-2). This was designed to enable participants to type a search name for an attribute within a tab rather than manually scrolling through the hierarchy. It was hoped that users without a medical background would now be able to find an attribute without needing to know where it might be in the hierarchy.



**Figure 4-2: New dimension tabs for round 2 with added search option.**

In round 2, the usability findings did not suggest a need to change the dimension tabs. The new search option proved very useful as participants used it to find attributes within the dimension hierarchy. In many cases, participants typed in part of the phrase (such as “labor and delivery” when searching for “Complications occurring mainly in the course of labor and delivery” or even part of a word in a phrase (such as “malignant neoplasm of pancreas” when searching for “malignant neoplasm of pancreas”). Stemming or partial phrases are supported by the searching algorithm and thus the participants had no trouble if the entire diagnosis name was not typed.

In round 3, it was discovered that while the new search box was helpful at times, it also caused confusion. The participants did not realize that when searching for a diagnosis, the diagnosis tab must be selected before the diagnosis search is submitted. In this type of example, even after correct spelling of the diagnosis name, the participant received an “Attribute not

found” message, leading to confusion and frustration during some of the tasks. In order to improve the search option of the dimension tabs, the name of the dimension was added to the search button (Figure 4-3). Thus, whenever a tab became active, the text on the search button changed to reflect the new dimension.

As evident from the figure, significant changes were made beyond the search box. The usability data from round 3 (the first round with the professionals) suggested that the participants had trouble formulating queries. The use of the dimension tabs to support the ad hoc nature of dragging and dropping attribute elements seemed confusing to the participants. It was decided to examine more closely the process of problem solving with SOVAT. Understanding the process of solving numerical-spatial problems for community health analysis is necessary for building an effective and usable decision support system. It was believed that at this point, SOVAT did not adequately support the process of solving numerical-spatial problems. Participants were constantly confused with how to set up a query and what essential components were needed to produce the query. A brainstorming session identified the numerical-spatial problem solving process to consist of:

- A numerical measure such as a death rate, cancer rate, or inpatient/outpatient hospital rate
- A temporal component such as a year
- A geographical component such as counties, municipalities, or custom-made communities.
- Additional data on which to filter the query such as: an age range, a particular diagnosis, or a particular sex or race.

It was believed that modifying SOVAT to support this numerical-spatial problem solving process was necessary for constructing a usable system. In supporting this process, the interface

would need to resemble more of a step-by-step systematic flow rather than a muddled multidirectional drag and drop environment seen in traditional OLAP environments. While this would constitute a massive overhaul of the interface's layout, it was felt it was a modification that needed to be made.

As can be seen in Figure 4-3, the numerical measure is easily selected by opening the drop-down list and selecting the measure with the mouse. In the previous versions, this was equivalent to opening the 'measures' tab and highlighting the measure. The selection of the temporal component (or year) is very similar. Instead of a drop-down list, the user has a standard list with which to select (using the mouse) the year. In the previous versions, this was equivalent to opening the year tab and highlighting the year. The filters section of the interface is not as straightforward. Filters are considered any dimension in SOVAT that does not represent a core component to a numerical-spatial routine; in this case, age, birth weight, education, race, region, and sex. These items will be used to filter the data, that is, given the components of a numerical-spatial query, these are the items to filter the result set. To add a filter item to a query, the user must select the attribute they need and hit the right arrow button to move it over to the 'filter list'. Any attribute in the filter list will be used as a filter on the data. Thus the more filter items, the more specific the result. If an item is brought over mistakenly, it can be removed from the filter list by highlighting it and hitting the left arrow button. This use of arrows to add and subtract items to a list is fairly well-known and seen in many software applications. In previous versions, the selection of filtering data (which was not referred to as that) was equivalent to opening one of these dimension tabs, and selecting the attributes needed.

In this version, it is not necessary to include any filters when performing a numerical-spatial ad-hoc query (the other possible query type, known as a ‘special query’, will be addressed at a later point in this chapter).



**Figure 4-3: Dimension tabs for round 4 of the usability study.**

The usability data from round 4 and 5 suggested that the different layout of the dimension tabs facilitated ad hoc query construction. Participants had little trouble with the lists for the year and the measure. Most of them understood the purpose of the filter tabs and were able to use the arrow to populate the filter list. In fact, one of the participants in round 4 indicated that the filter tabs were one of the best aspects of the system, responding that the “filters were quite easy to use”.

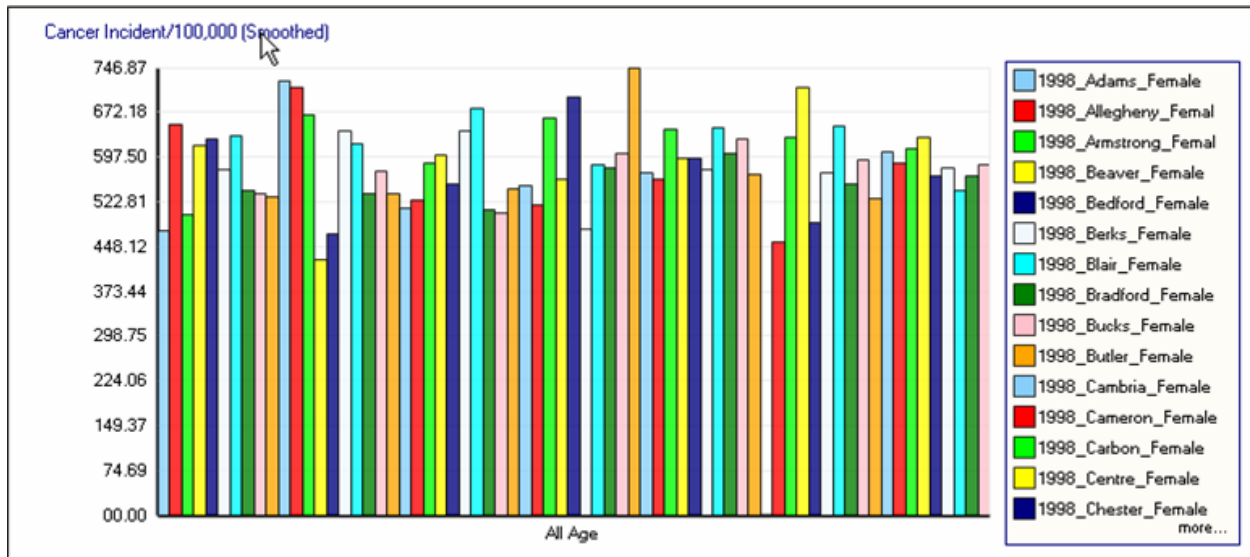
In both round 4 and 5, the search box proved to be very helpful in finding attributes. The participants had little trouble with this feature. The addition of the dimension name on the

search button seemed to eliminate any confusion related to which dimension would be searched. In fact one participant in round 4 indicated that one of the best things about the system was that by using the search option it was “easy to find ICD [diagnosis] and [other] groups”. The usability data from rounds 4 or 5 did not suggest that changes needed to be made to the search box or the dimension tabs.

#### **4.5.1.2. Chart Display**

In round 1, the participants used the chart mostly as a data presentation feature. They would look to the chart to interpret the results of their query. Direct manipulation with the chart, such as right clicking on a chart element (such as a bar) and drilling down, was rarely used. One area of confusion was the interpretation of the chart (Figure 4-4). The chart does not have an English language title that clearly describes the query. In fact, the only label above the chart is the numerical measure in the query. This feedback seemed to be insufficient to the participants. In many instances the participants were frustrated as they struggled to interpret the chart. In order to enhance the usability of the chart, a title was added on the top to provide an explanation of the result.

An additional area of confusion was with the legend on the chart. On instances in which the space of the legend could not include all of the chart elements, a “...more” label was displayed at the bottom of the legend. This gave the participants an indication that pressing the label would show them more elements in the legend. However, the label is simply text and clicking on it does not produce any function. This provided frustration to the users as they continued to click on the label without getting an interface response. Unfortunately, this chart feature was provided by the 3<sup>rd</sup> party software developer who supplied the chart layer and can not be altered in any fashion.



**Figure 4-4: SOVAT chart for round 1 of the usability study.**

In round 2, it was apparent that the title was very helpful in interpreting the chart. Instead of just the numerical measure, the title now included the attributes within the query (Figure 4-5). The format consists of the measure first and then the other attributes separated with the word ‘by’ such as “Population by (Geography=Adams, Allegheny, Armstrong...), by (Year=2000) by (Age=All Age).” Ideally a true English language phrase (such as “2000 Pennsylvania Population by County”) would have been used, however it was felt this would be too time consuming to develop and thus it was decided not to implement it. Despite a true English phrase, the new title was well perceived. In fact participants listed “chart display” as one of the best aspects of the system in round 2.

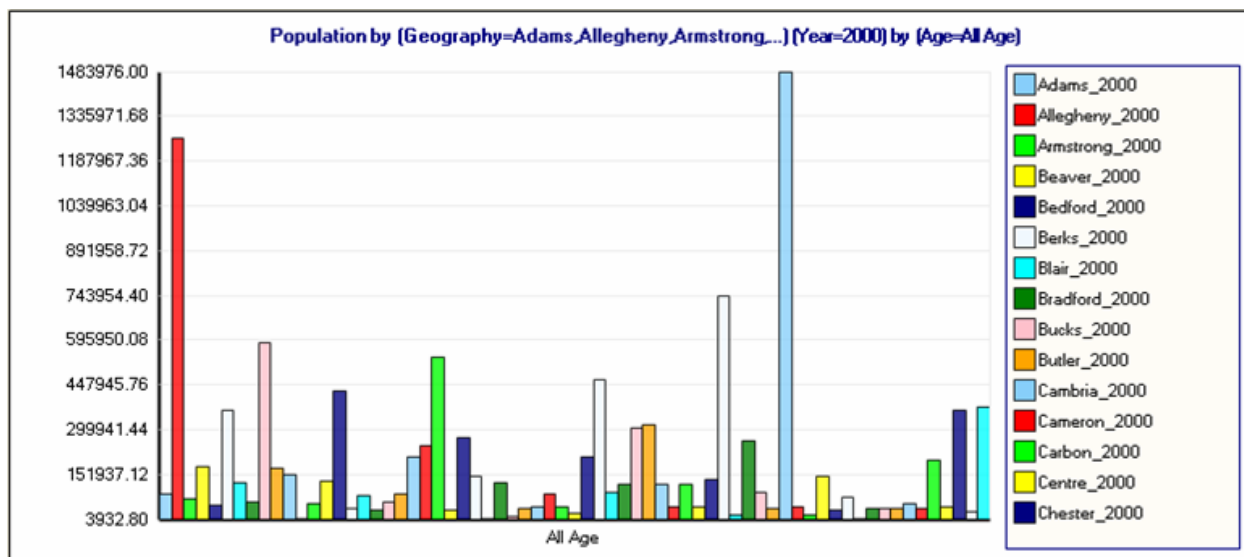
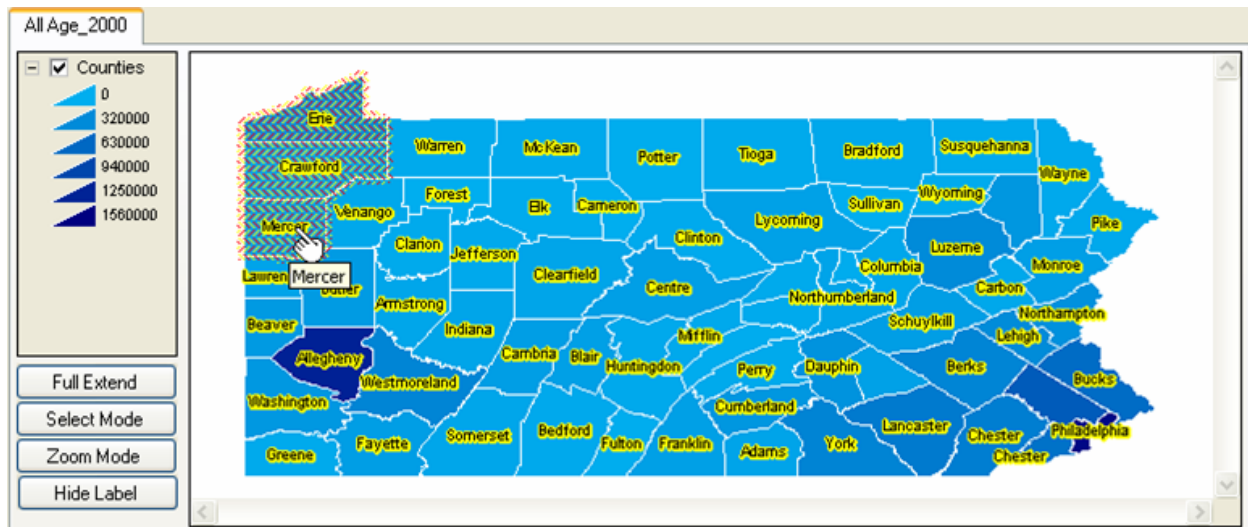


Figure 4-5: SOVAT chart for round 2 of the usability study.

In round 3, 4, or 5 usability findings suggested that no modification to the chart was necessary. The only change to the chart at all was the addition of the numerical measure on the Y-axis after round 3. This was not derived from the usability findings, but rather the desire to make the chart look a little more like a “traditional bar chart” found in the literature.

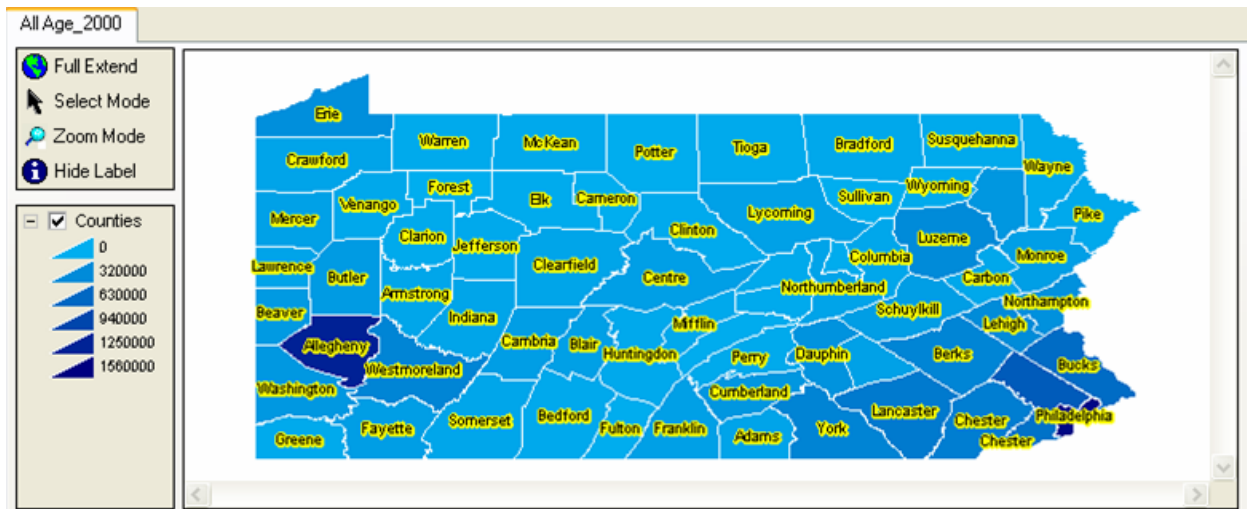
#### 4.5.1.3. Map Display

In round 1, the majority of the erroneous actions related to attempts of direct manipulation with the map. This occurred when participants needed to produce a query with geographical elements and attempted to drag and drop map elements (selected counties or municipalities) onto the charting area. Drag and drop actions could not be carried out with the map elements. Based on this limitation, the participants were not certain how to compare the current selection against the other area(s). It was felt that enhancing the tools for direct manipulation of the map through drag and drop operations would facilitate this process. Users would then be able to drag one set of geographical areas and then repeat this action to perform the comparison (Figure 4-6). Allowing for map elements to be dragged and dropped might enhance system usability for this type of sub-task and reduce the number of problem space and erroneous actions.



**Figure 4-6: SOVAT map for round 2 of the usability study.**

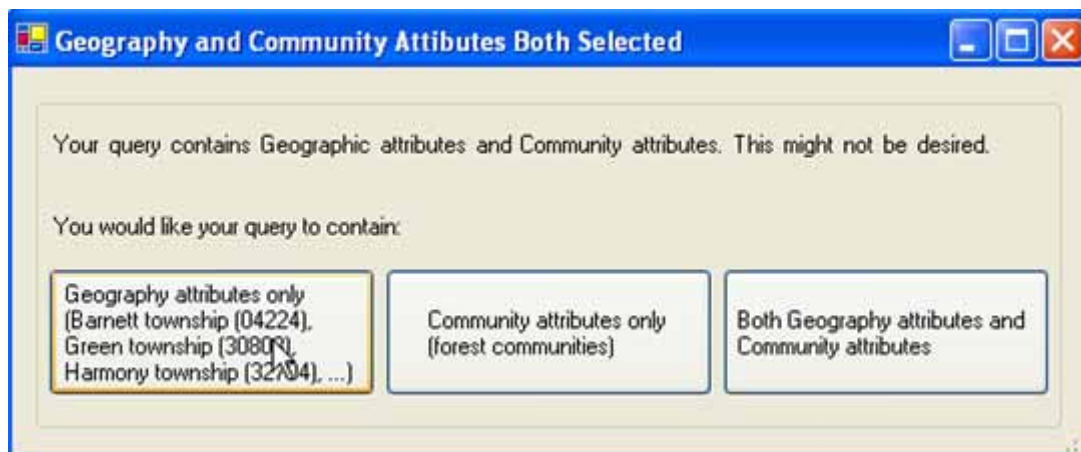
While use of the map was facilitated through the addition of the drag and drop querying option, round 2 indicated the map contained more usability issues that needed to be addressed. Users did not seem to understand the concept of drilling down and thus often seemed ‘lost’ after performing this action on the map. They were not certain where they were after drill-down from a county and whether a query or a simple ‘zoom’ was actually performed. In addition, map selection created problems as users either had difficulty selecting a geographic area, or accidentally drilled down by double-clicking on a county. This added significant frustration. It was decided to change the map viewing tools such as zoom, select, full extend, and hide label from buttons to icons (Figure 4-7). It was believed that this would better convey their function and purpose on the interface as simple map viewing tools rather than querying functions (like drill-down).



**Figure 4-7: SOVAT map for round 3 of the usability study.**

Another usability issue was related to an instance when a participant submitted a query with a custom-made community. The community attributes fall under a different dimension tab than the geography attributes. However, they are both geographic-defined dimensions and thus the participant would need to deselect any items currently selected on the geography tab before submitting a query with the community. If the geography attributes are not de-selected then they would be included in the query. This would cause frustration to the participant as he/she would wonder why the map contained color-coded areas that were not included in the community. The worst-case scenario is when all counties are selected on the geography tab (which is the case at startup). Failure to deselect these attributes when querying with the community attribute will ‘cover up’ the community on the map. This is because all the counties on the map will be color coded since they are all included in the query. Thus, the community would be very difficult to visualize. This occurred in several instances and caused significant frustration for the participants. It was decided that in order to enhance the usability of the map, an effective feedback message would be created that would prompt the participant to decide whether to

include geography items when querying with custom-made communities. The feedback message for this problem is shown in Figure 4-8.



**Figure 4-8: Help message when a community-related query is submitted.**

The feedback message pops up when the user submits a query containing selected items from both the geography and community dimensions. This message helps them decide what to include, while the system automatically de-selects the attributes from the undesired dimension (unless the ‘Both’ selection was chosen).

In round 3, the new icons seemed to clarify their purpose as non-querying tools. However new usability issues were discovered after reviewing the think-aloud data. Some of the participants in this round were unfamiliar where specific counties lie in Pennsylvania. This caused them significant time as they needed to search (almost ‘row by row’) the map for the county of interest. This caused frustration, especially when it took more than one pass to find the county. In order to enhance the usability of the map, it was decided to add a search box, much like the one found on the dimension tabs (Figure 4-9). This would allow the participants to type in the name of the county or municipality and hit the search button. The area would then be highlighted on the map for the user to see.

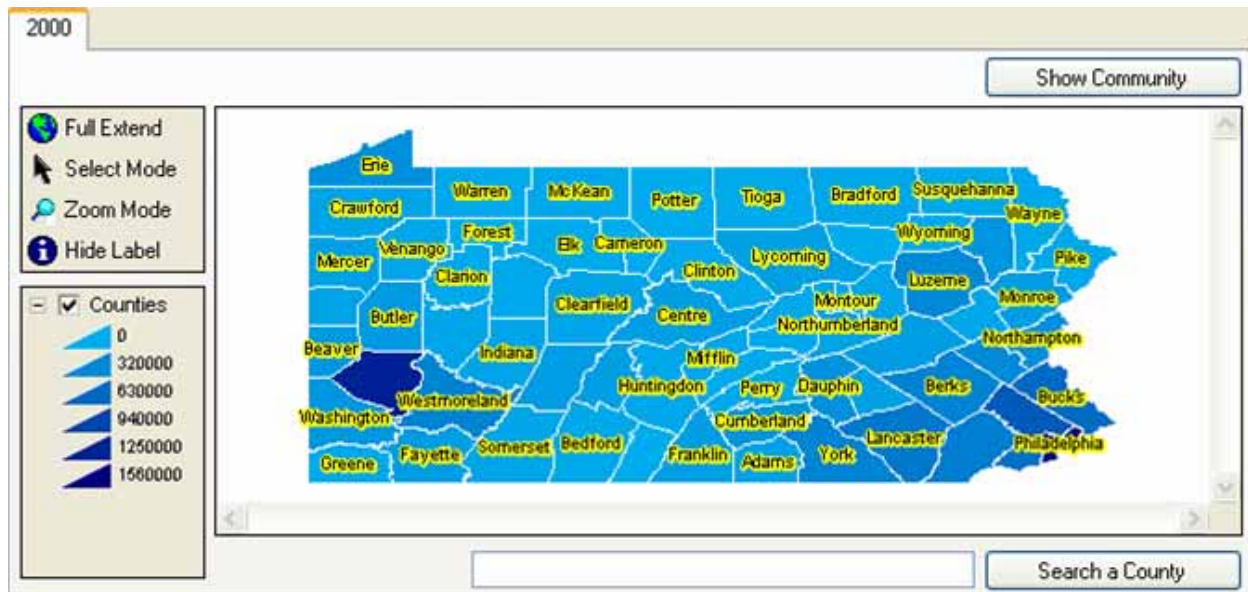


Figure 4-9: SOVAT map used in round 4 of the usability analysis.

Another area of confusion for the users was the meaning of the custom-made communities. Many users were uncertain how this related to the use of map and other geographical items such as an individual county or municipality. In order to better represent their purpose, the community list was taken from the dimension tab area and added to the map portion of the interface (Figure 4-9). It was anticipated that being included in the map section of the interface would clarify the fact that custom-made communities are considered geographical attributes.

Finally, the feedback message improved usability issues with the map. Participants no longer were faced with the difficulty of interpreting a map that contained overlapping geography and community color-coded areas.

The big usability issue in round 4 was that many participants became frustrated when they accidentally drilled down when they only meant to select a map item. Any occurrence of a double-click would produce an automatic drill down. This caused considerable frustration as the participants didn't realize what was happening and why they all of a sudden were looking at a different map. This forced them to drill back up and start their map selection process all over

again. In order to enhance the usability of the map, drill-down would now be performed by a right-click rather than a double-click. This would hopefully reduce the incidence of unintentional drill-downs by requiring the user to right-click and then select drill-down from the menu. Double-clicking on the map now would only reselect the map item.

It turned out that the placement of the communities tab was not optimal. When the tab was opened it covered a portion of the map. During one of the tasks (task 2) the participants were required to select a portion of the map that became covered if the communities tab was opened. This created significant frustration as they had to figure out what happened to the counties and how to close the communities list. One participant was significantly frustrated as it took over 5 minutes to close the list and uncover the counties. It was decided that the community list should be moved so that it would not cover up any portion of the map when it was opened. The final version of the map is shown in Figure 4-10. The community list was moved from the top left corner of the map, to the side of the map by the legend. This would ensure that the list would not cover up any portions of the map.

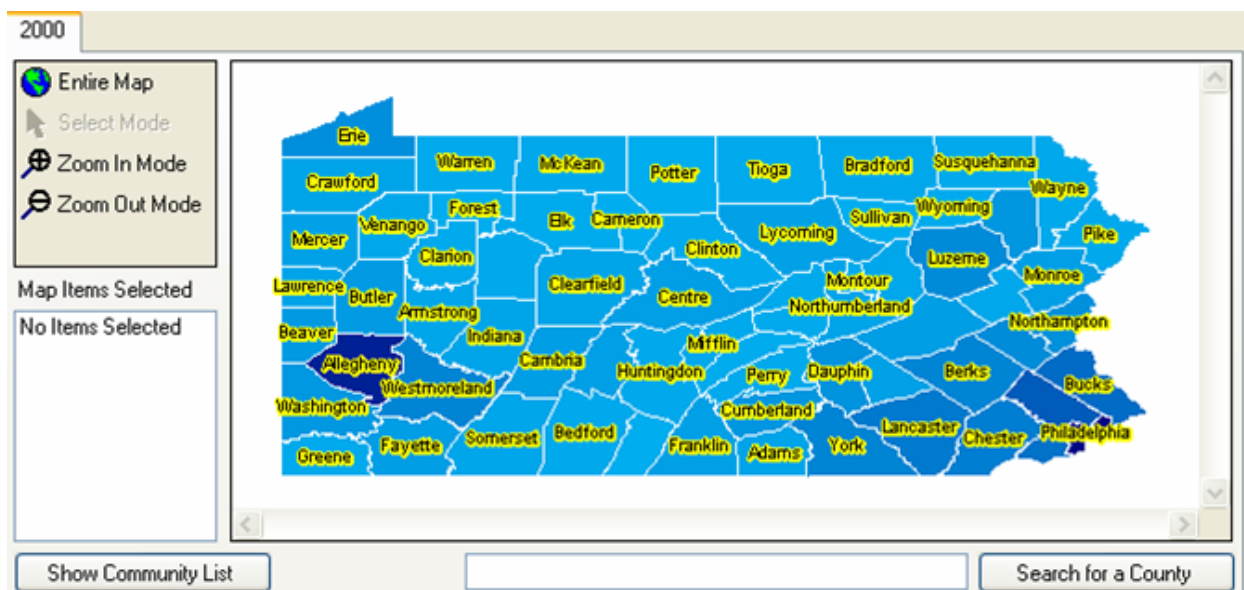


Figure 4-10: SOVAT map for round 5 of the usability study.

Although rarely used, it was decided that the zooming functions needed to be facilitated. Zooming is important especially when viewing the municipality map. Some of the municipalities are very small and hard to view at the normal viewing mode. During these circumstances, a zooming tool would be very useful in order to view and select these elements. The current actions needed for zooming were very confusing. In similar fashion to many Windows applications, the participant would need to select “zoom” and then go to the map and draw a clockwise circle around the area in which they wanted to zoom. A similar method needed to be done for zooming out. Through the usability findings, it was apparent that the participants would never grasp this concept. Zooming would likely be a more usable function if it were a click action instead of a circle action. It was decided to modify zooming so that the participant would select whether they wanted to zoom in or zoom out, and then perform a regular mouse click on the map where they wanted this action to occur.

In order to enhance usability related to selection of map elements, a list box was added to the final version of map in order to confirm to the participant what specific items were currently selected on the map. Some of the usability data suggested that the participants were not always certain which counties or municipalities were selected. This would cause errors with the answers since a county might be omitted from the results. Introducing a list would hopefully confirm that all the necessary counties or municipalities were in fact selected. The list was added in the location where the map legend used to be. Space was a factor in deciding where to add additional interface items. The usability data suggested that the map legend was of little use. The participants recognized the concept that darker colors on the map indicated a higher numerical value than lighter colors. If participants needed to see actual numerical values, they

referred to the chart. Thus in order to make room for the new list, the legend was removed from the interface.

Round 5 suggested that the modifications made to the map enhanced its usability, especially in relation to direct manipulation. Requiring the participants to right-click instead of double-click when drilling down eliminated any instances of this occurring by mistake (such as when the participant simply needed to select a map item). Map selection was also facilitated by the “map items selected” list. The findings suggest that the new placement of the communities list reduced usability problems related to its interaction with the map. The list now opened to the left of the “map items selected” list and thus did not obstruct any interface component. The new zooming tools definitely improved the usability of this mapping feature. This helped the participants especially in task 5 which required viewing of small municipalities. The participants did not have any trouble using the zooming feature to perform this action.

With all these changes, the final round did not suggest that any map component needed to be modified.

#### **4.5.1.4. Row/Column, and Special Functions**

In round 1, the use of the row and column feature was found to be a very confusing concept. There were many examples that demonstrated that the participants were not clear of its purpose. For example, they frequently tried to drag and drop an attribute from a dimension tab into either the row or column list instead of dragging and dropping a dimension name from the background list (Figure 4-11). Thus, they failed to recognize that the concept of “row and column” is dimension specific and not attribute specific. Once they did recognize this notion, the tasks requiring Top 5 or ranking analysis put significant requirements on them to place the appropriate

dimensions in these lists. This produced many erroneous actions and problem space issues as incorrect dimensions were frequently chosen.

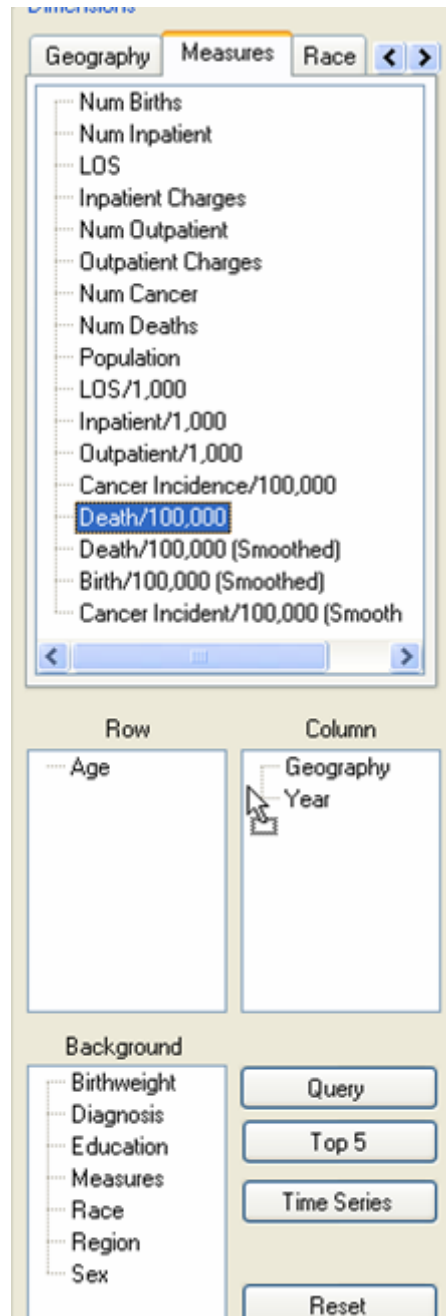


Figure 4-11: Row/Column error from round 1 of the usability study.

Based on the difficulty that the row/column concept presented, it was decided to implement a wizard to facilitate this process. Instead of asking the participant to drag and drop dimension

names to the appropriate ‘row’ and ‘column’ for producing a Top 5 query, SOVAT would have a ‘Top 5 Wizard’ that would use easily understood English sentences to help the user construct a proper Top 5 solution. For example, if the task is to find the “Top 5 cancer diagnosis in McKean County for males in year 1999”, the first question in the wizard could be:

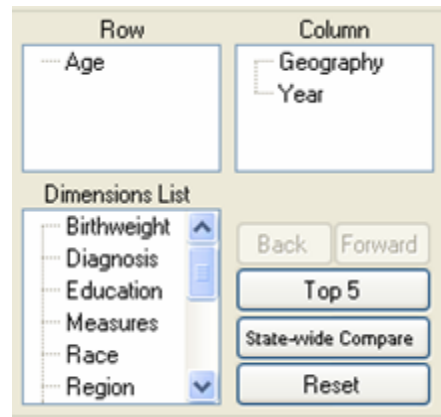
What would you like to see the Top 5 of?

- Counties in the State
- Municipalities in a County
- A type of Diagnosis

Further similar English sentence questions would be used to get the other variables. This was perceived as a much easier way of constructing a Top 5 task and places no onus on the user to understand the row and column concept for these types of queries. It was hoped that this would “mask” the selection of row and column dimensions and allow for ranking analysis to be easily constructed. Figure 4-12 shows the implemented wizard.



**Figure 4-12: Top 5 Wizard implemented in round 2 of the usability study.**



**Figure 4-13: Row/Column and special features component for round 2 of the usability study.**

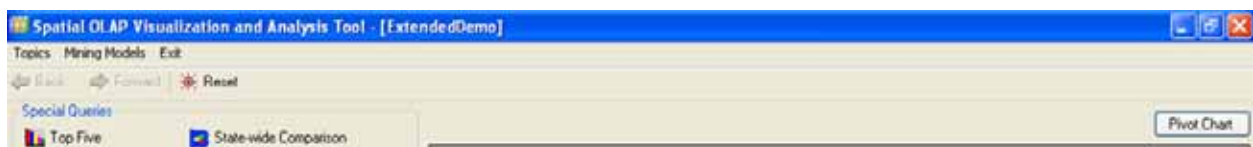
The modification for round 2 had minimal success. The top 5 and state-wide comparison wizards were effective when used, but the participants in many instances did not ‘notice’ them. This might have been the fact that they looked like regular buttons rather than special icons. It was then believed that single action processes should be shown as a button, while multi-step processes should receive an icon (Figure 4-14). This would hopefully increase their utilization. The tasks asking for top 5 and state-wide comparison still received many erroneous actions; however, it was believed that this was caused by the inability to notice the wizards on the interface rather than anything else. Thus making them more accessible would hopefully reduce the number of errors associated with these tasks. There were no direct modifications to the row and column concept; however they contributed to fewer erroneous action and problem space instances during this round. This might have been due to the fact that for the most part, they were left alone. Participants in this round focused much more on the comparison of geographic elements than the use of the rows and columns.



**Figure 4-14: Special Function components for round 3 of the usability study.**

In round 3, the row/column concept proved to be extremely frustrating for the participants. It led to many errors, confusion, and increase in time to completion. It was clear from the think-aloud data that drastic changes needed to be done with this interface feature. The participants did not seem to grasp the purpose of the row and column. The special function icons did not fix the usability problems with the row and column as was originally thought. These functions were helpful during specific tasks; however row/column was still an interface feature that the participants attempted to utilize for other tasks. When this utilization occurred, problems ensued, such as: allocation of improper dimensions to either the row or column list, failure to include a single dimension in both the row or column list, or use of the dimension names in either the row or column list as a drag and droppable querying element. It was then concluded that the presence of row/column on the interface did not correspond to its importance as a necessary feature for numerical-spatial problem-solving. It was more of a data orientation feature than anything else. Its purpose is to affect the presentation of the data on the chart; changing a dimension from a “row” dimension to a “column” dimension would switch it from the legend to the x-axis on the chart. As such, the magnitude of space it contained on the interface as well as its form as a drag and droppable feature seemed inappropriate given its true importance in problem solving. In order to enhance usability of the interface, it was decided row/column would be removed from the interface. In its place, it was decided to add a “pivot” button on the chart section of the interface. For the most part, this would fulfill the purpose of row/column while hopefully eliminating the usability issues it caused. In addition, the placement of the pivot button in the chart area of the interface seemed appropriate since it only affected the presentation of the chart and not the map.

The usability data suggested that the special functions needed to be represented on the interface as a distinct querying component. That is, the icons needed to be grouped as “special queries”. The dimension tabs, as previously discussed, were represented as components for ad-hoc query formulation. This would hopefully convey that they are different methods of performing queries. This modification can be seen in Figure 4-15. The “pivot chart” button is also included in this figure to show its new location on the top right corner of the chart area.



**Figure 4-15: Special Function components for round 4 of the usability study.**

The usability data from round 4 suggested the changes to row/column and special query features enhanced the usability of the interface. Users were now able to clearly distinguish between special function icons and the rest of the components used for ad hoc query formulation. The data showed a frequency in use of the special queries. Once “inside” the wizards, the participants had little trouble completing the queries. The pivot chart button eliminated any errors from previous rounds that involved interaction with the row/column component. The participants seemed to understand that the button was used to simply orient the data on the chart. The participants that used it clicked the button once to see the result of the action, and then almost always immediately clicked it again to return the chart to its original orientation. The success of the modifications was evident in that the usability data from rounds 4 and 5 did not suggest additional changes to the special queries or pivot chart needed to be made.

The final version of the SOVAT interface with all of the discussed modifications is shown in Figure 4-16.

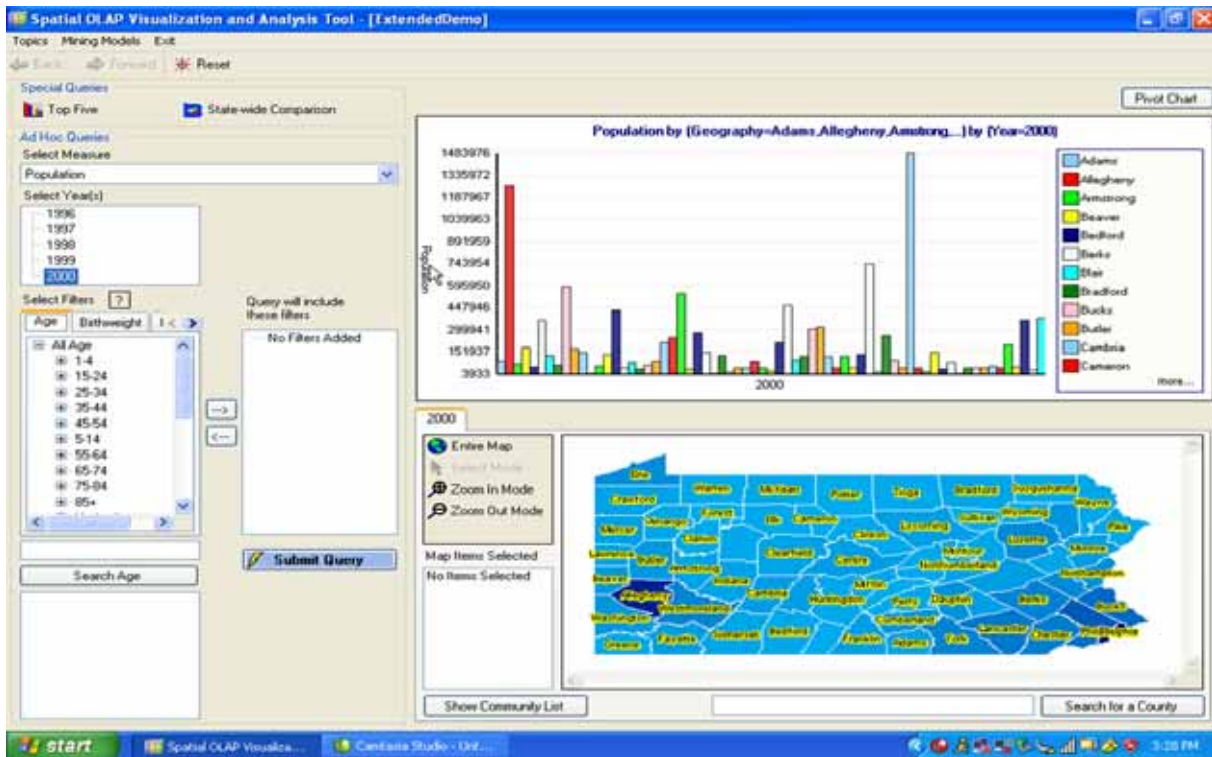


Figure 4-16: Final version of SOVAT interface after round 5 of the usability study.

#### 4.5.2. Best and Worst Aspects of the Interface

As indicated previously, the participants were asked to complete 2 post-study questionnaires. One of these was more open-ended that asked them to list the three best and worst aspects of the interface (**Appendix G**). This form proved very helpful in evaluating the usability of SOVAT after each round. When analyzing the findings from the first and last rounds (round 1 and round 5), some interesting patterns can be seen that indicate the system is much more usable (see **Appendix H** for the best and worst findings from rounds 2-4).

In round 1, 7 total “worst things” were noted from the three participants (Table 4-1). Four of these 7 aspects related to “support”, such as: “there were few directions”, “could have been given a demo session”, “troubleshooting”, and “no help menu”. In round 5, 7 total “worst things” were noted from the participants (Table 4-1). In the responses this time, there was no mention of these “support” issues. The lack of reliance on directions, a demo, a help menu, or

troubleshooting components, suggests that the system is much more usable. By round 5, the interface did not require the participants to rely on support components to complete the tasks. The issues mentioned in round 5 seemed much minor in severity than round 1. These (5 out of 7) were mostly related to difficulty in viewing the map at the municipality level. This is a usability issue that needs to be addressed; however it is far less significant than the need for support items (in round 1).

**Table 4-1: Worst aspects of the SOVAT system from round 1 and round 5.**

	<b>Round 1</b>	<b>Round 5</b>
<b>Worst Aspects</b>	<ol style="list-style-type: none"> <li>1. There were few directions</li> <li>2. I could not go back to the previous task.</li> <li>3. I could have been given a demo session.</li> <li>4. Not enough time</li> <li>5. Troubleshooting</li> <li>6. No 'Help' menu</li> <li>7. No 'Search' option</li> </ol>	<ol style="list-style-type: none"> <li>1. The maps are small. We need to easily find different levels of maps</li> <li>2. We need a history for what we did</li> <li>3. Not easily find the areas in the plot and connect them into the map</li> <li>4. Hard to find municipalities on the state map.</li> <li>5. Inability to press enter to perform a search (had to press mouse instead)</li> <li>6. Zoom in/out feature was a little confusing</li> <li>7. Small print when in municipalities map (hard to select/de-select)</li> </ol>

Examining the “best aspects” also shows patterns in improvement of the usability of SOVAT. Table 4-2 shows the best aspects listed by the participants from round 1. Here, 6 of the 9 comments are related to the appearance of the interface such as: “the look of the interface, the visualization of the data, the design is clear and beautiful”. These are important characteristics to have in a system; however an interface that looks good on the screen isn’t necessarily usable and valuable for problem solving.

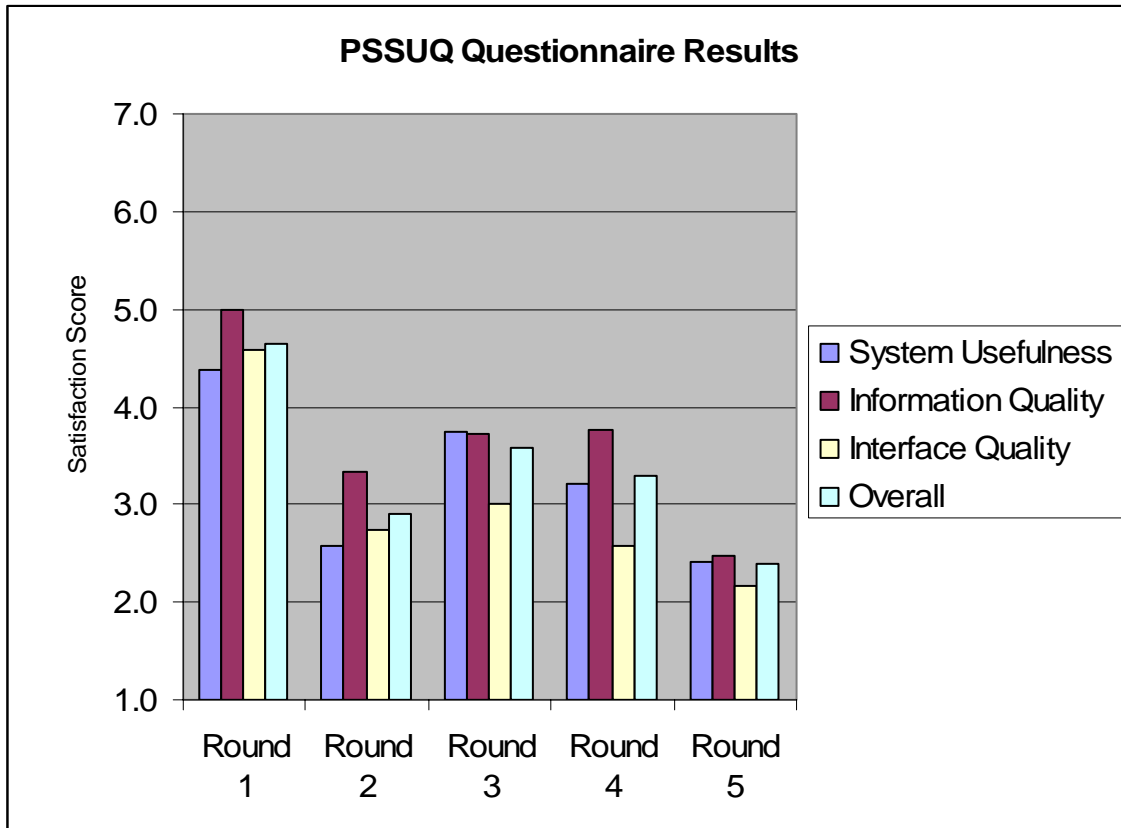
**Table 4-2: Best aspects of the SOVAT system from round 1 and round 5.**

	<b>Round 1</b>	<b>Round 5</b>
<b>Best Aspects</b>	<ol style="list-style-type: none"><li>1. Display of data</li><li>2. Easy to see medical information</li><li>3. Helpful for spatial analysis</li><li>4. Look [of the interface]</li><li>5. Spatial and data [numerical] together</li><li>6. Visualization of data</li><li>7. Windows [interface] design is clear and beautiful</li><li>8. Covers very completely</li><li>9. Easy to use (with a little more training)</li></ol>	<ol style="list-style-type: none"><li>1. The category to choose is clear</li><li>2. Help system is ok</li><li>3. Easy-to-see layout</li><li>4. Very convenient to search for diagnoses</li><li>5. The self-directed Top 5 and state-wide questioning</li><li>6. Easy access to items used to create queries</li><li>7. The ease of quickly getting info</li><li>8. The search for diagnoses without picking out exact words.</li></ol>

In round 5 (Table 4-2), 7 of the 8 comments related to interacting with the system, while only 1 aspect related to the appearance of the interface. These comments are aspects such as, “The category to choose is clear”, “the help system is ok”, it is “very convenient to search for diagnoses”, and the “ease of quickly getting the info”. These comments suggests that participants view the system as not simply having an interface that looks good, but rather a system that is helpful, easy to use, and allows them to perform numerical-spatial queries.

#### **4.5.3. Subjective and Objective Results**

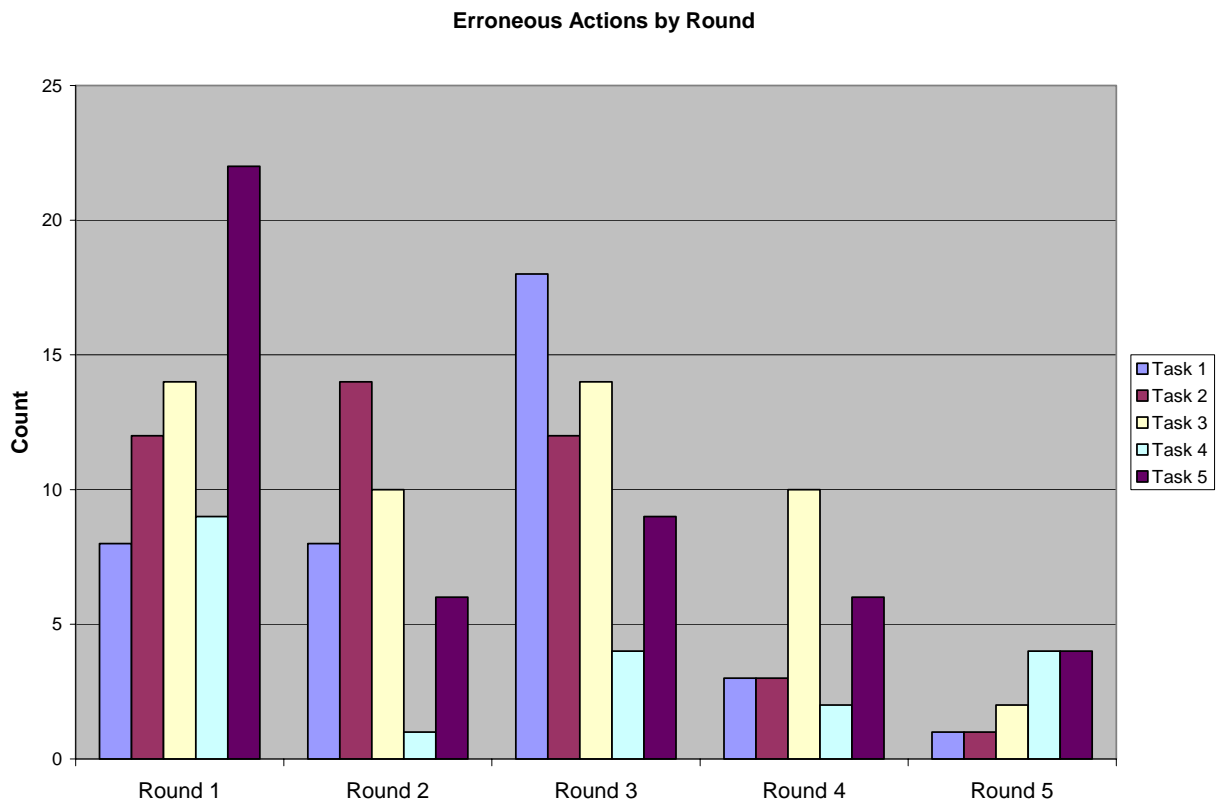
Figure 4-17 shows the subjective results from the Post-Study System Usability Questionnaire (PSSUQ) (see **Appendix I** to view the tabular representation). The results shown here are summarized by round. Lower numerical values indicate a higher level of user satisfaction. As mentioned, in addition to overall satisfaction score, the responses can be divided into three sections: system usefulness, information quality, and interface quality.



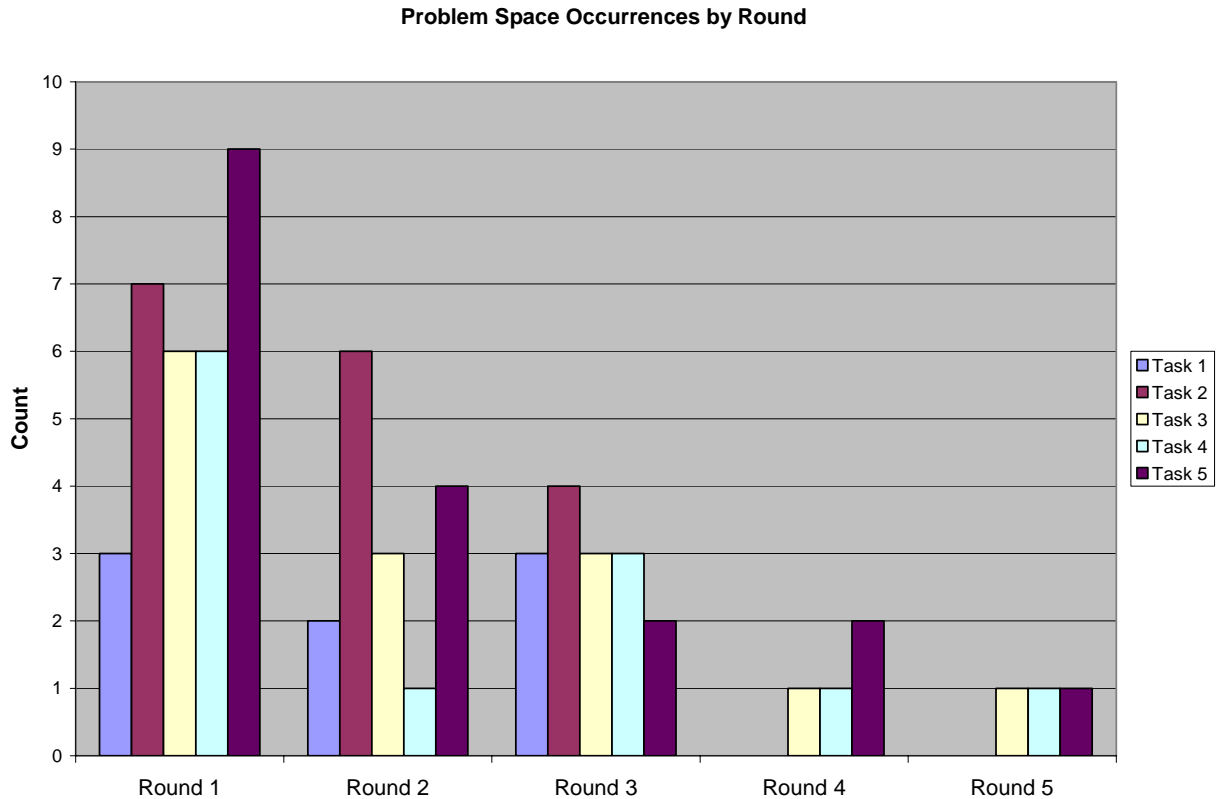
**Figure 4-17: PSSUQ results summarized by round.**

The data suggests improvement in all three usability categories (system usefulness, information quality, and interface quality) from the first round to the final round. Examining data between the rounds, the overall satisfaction score improved by 61% (from 4.65 to 2.89) for the graduate students between round 1 and 2 but got 45% worse with the first round of professionals in round 3 (2.58 to 3.75). Round 4 showed a slight improvement with a score of 3.28 (5%), while the final round, round 5, was clearly the best at 2.39 (a 37% improvement from the previous round). System usefulness, which measures the users' perception of how the system can improve their job performance, improved the most with a 70% change (4.38 to 2.58) between round 1 and 2, but was the worst from round 2 to round 3 with a 45% negative change (2.58 to 3.75). It improved over the final two rounds with a final-round score of 2.42 (a 33% improvement from round 4).

The objective summarized results suggest improvement in the system based on all four usability criteria; time, erroneous actions, problem space, and answer, from round 1 to round 5 (see **Appendix I** to view the tabular representation). The following two charts (Figure 4-18 and Figure 4-19) show the results for erroneous action and problem space occurrences respectively.



**Figure 4-18: Erroneous actions summarized by round.**



**Figure 4-19: Problem space occurrences summarized by round.**

The charts show a nice improvement after round 3. This suggests the drastic modifications after round 3 produced a more usable system. The changes were intended to aid the participant in construction of numerical-spatial queries. The data suggests that the participants were much more efficient in solving the tasks; in that they had fewer wasted/erroneous actions and did not need to try numerous methodologies in solving the tasks. This improvement is also seen by looking at the other objective measurements; time and answer (Tables 4-3 and 4-4). Time appears to improve a lot for the final round (round 5) except for tasks 4 and 5 which did not improve from round 4.

**Table 4-3: Time (in min) summarized by round.**

	Time (min)				
	Round 1	Round 2	Round 3	Round 4	Round 5
<i>Task 1</i>	TNC*	9	17	6	4
<i>Task 2</i>	TNC	14	16	9	5
<i>Task 3</i>	TNC	12	TNC	14	7
<i>Task 4</i>	TNC	8	TNC	12	12
<i>Task 5</i>	TNC	12	18	13	12

TNC ~ Task Not Completed

**Table 4-4: Answers summarized by round.**

	Answer				
	Round 1	Round 2	Round 3	Round 4	Round 5
<i>Task 1</i>	0 Correct	3 Correct	3 Correct	3 Correct	3 Correct
<i>Task 2</i>	1 Correct	2 Correct	3 Correct	2 Correct	3 Correct
<i>Task 3</i>	0 Correct	0 Correct	1 Correct	2 Correct	2 Correct
<i>Task 4</i>	0 Correct	3 Correct	2 Correct	3 Correct	3 Correct
<i>Task 5</i>	1 Correct	3 Correct	2 Correct	3 Correct	2 Correct

Table 4-5 shows the head-to-head results across all the rounds. A round received a ‘win’ if it had the best result when compared to the other rounds. In the event of a tie, the rounds with the equal results got the win. Since there are 20 total results (5 tasks x 4 types of objective criteria), the best possible score is a 20, while the worst possible score is 0. The last round, round 5, was clearly the best round as it received a total of 17 wins. The only places it did not get a win were time and erroneous actions in task 4, and answer in task 5.

**Table 4-5: The number of wins when comparing each round.**

	Round 1	Round 2	Round 3	Round 4	Round 5
Wins	0	7	2	8	17

#### 4.6. Discussion

Due to the rigorous usability study, the SOVAT interface changed dramatically from the beginning of the study. The results from round 5 indicate that a usable system has been

developed. An important and informative final step in this study is the discussion of visual design criteria of ‘before’ and ‘after’ versions of the interface. This enables for discussion from a HCI standpoint, how the interface has changed. In addition, it highlights which visual design aspects have been addressed and which ones have not been addressed. There are many publications that list visual design components, however, for the purposes of this write-up, the principles used will be based on Mullet and Sano’s “Designing Visual Interfaces”[70]. The 5 principles they list for designing effective interfaces are: *elegance and simplicity*, *scale*, *contrast and proportion*, *organization and visual structure*, and finally *module and program*. These principles will not be covered extensively, yet they will each be addressed briefly in order to analyze how the interface changes can be applied to recognized visual design concepts. Comparison for these purposes will be based on the initial SOVAT interface before round 1 (Figure 4-1) and the interface used in the final round (Figure 4-16).

*Elegance and Simplicity.* This visual interface principle refers to carefully selecting the interface elements and ensuring they are presented to the user in a well-designed and straightforward manner [70]. The lower left corner of the original interface appears to not embody this principle. There is significant clutter as the row, column, and background list are meshed together with wizard, time-series, and reset button; which have little to do with these lists. This region created significant confusion to the user in these early rounds. The latest interface has ‘cleaned up’ this area by removing the row, column, and background list. In addition, the nearby wizards that had little relation to these lists have been moved to their own section of the interface. The result has seemingly created a less complex and cluttered interface for the user.

*Scale, Contrast, and Proportion.* This visual interface principle refers to appropriateness of the clarity of the display among the different interface components [70]. This can refer to size (something too small or too big), color (something too light or too dark), or appearance (something too prominent or too indistinct) [70]. The most significant issue related to this theme is the display of the municipality-level map. In the early versions, the lack of an effective zooming feature made this interface component extremely difficult to interpret. Participants had trouble identifying and selecting specific municipalities. The addition of a more straightforward zoom-in/zoom-out feature now allows the users to more easily interpret the initially small geographic areas on the map. The zoom-in feature enables the size of the municipalities to increase significantly until an appropriate viewing scale is reached. While the users did note in the final round that the interface could improve with this issue, it is evident that significant strides have been made.

*Organization and Visual Structure:* This visual design principles refers to the need to provide the user with an organized and systematic structure that facilitates system interaction [70]. It is this area, where the interface has improved the most. The early interfaces supported a drag and drop ‘a la carte’ manner in which to create a query. Participants did this by accessing the dimension tabs in the upper left corner of the interface. There was no order to this process. The participants could pick and choose what they wanted without having any direction as to what to do next. This chaotic process was very daunting to the participants at times.

The latest interface addressed this by better organizing interface elements. The interface takes on a more systematic flow from the top toolbar with navigation functions, to the query forms, and finally to the data display of the map and the chart. In addition the flow within an ad-hoc query from measure → year → filters → geography seemingly makes the interface much

more elegant and simplistic in such that order has replaced the chaos from the earlier versions. This structure directs the decision making in a topological fashion: *What type of query do I need (special query or ad-hoc query)?; If ad-hoc, what measure should I select? I selected the measure, so now, what year do I need?; The measure and year have been selected, do I now need any filters?; I have everything organized, finally, what are the geographic elements that I would like in my ad-hoc query?*

*Module and Program:* This final visual design principle refers to the importance of providing regularity and structure to the interface through the development of the individual interface components (modules) [70]. The SOVAT interface from the beginning consisted of three different modular components: Microsoft Windows controls, chart presentation, and GIS presentation. These multiple components presented significant difficulties from a visual design perspective. Some of these issues are not able to be fixed. For example, there was confusion throughout the study about the difference in color between the bars in the bar chart and the color-gradated map. Coordinating these two modules so the color is consistent will be very difficult. The most positive interface development for this principle is related to the special query wizards (Top 5 and state-wide comparison). It is important, as the user iterates through the steps of either wizard, to maintain consistency and uniformity from one page to another. That is, besides the content of each form, the layout should be identical; what the user sees from a visual perspective on the first form should be consistent with what is seen on the final form. This is the case both within each wizard and across each wizard (as the forms used in both are seemingly identical to the user). Within these forms, the search options are in the same place, they have the same size and they work in the same fashion. In addition, each form has a 'Next', 'Back', 'Finish', and 'Close' option which is all activated based on the presence or absence of user selection. Upon

pressing 'Finish', the user sees the result clearly displayed in both the chart and the map as if it was as an ad-hoc query. This certainly creates a sense of uniformity and consistency across the different modules within the system.

#### **4.7. Limitations**

It was understood that participants who were not representative users were not ideal participants. The major reason for not using only representative users was because of the value of their time (away from work) as well as the difficulty of recruiting a sufficient number of them to fill every phase of the evaluation. It is believed that using typical future users such as public health graduate students yielded similar findings.

#### **4.8. Conclusion**

This chapter described the rigorous usability study of the SOVAT interface for numerical-spatial problem solving during community health assessment analysis. In total 5 rounds were used with 3 individuals per round (for a total of 15 people). Both future users (public health graduate students) and representative users (community health assessment professionals) participated in the study. Evaluation and system enhancement occurred after each round. Think-aloud protocol was used as the participants completed 5 community health problem solving tasks. Screen capture software and a microphone were used to record their interaction with SOVAT. After completion of the tasks, the participants were asked to complete two post-study questionnaires aimed at identifying their satisfaction with the system. Objective and subjective measurements were used to identify usability issues. Objective measurements consisted of: time to task completion, erroneous actions during each task, problem space or number of different

methodologies used to solve the problem, and finally whether the verbal answer they gave was correct. Subjective measurements consisted of the responses from the post-study questionnaires.

The objective and subjective results showed a significant increase in usability and user satisfaction with SOVAT. Objectively, the improvements made to the system reduced the time of task completion, the number of erroneous actions, the number of problem space occurrences, and increased the accuracy of the responses. Subjectively, the system improved in all areas including: perceived usefulness, information quality, and interface quality.

The most significant change to the interface during this evaluation occurred after round 3. The system was altered to better reflect the systematic process of numerical-spatial problem solving including the separation between special and ad-hoc queries and the incorporation of optional filter tabs. From a visual design perspective, the interface has evolved from a chaotic un-orderly presentation, to a more fluid and elegant presentation of numerical and spatial themes.

With the development of an effective and usable system, SOVAT can now be introduced as a decision support tool for community health analysis. Before a summative evaluation study can begin, the current technologies used for this type of decision support must be identified. This will enable the selection of a proper comparison for the final evaluation. Attention will now be turned to the community health assessment survey aimed at addressing this issue.

## **5. INFORMATION TECHNOLOGY UTILIZATION IN COMMUNITY HEALTH ASSESSMENT ANALYSIS**

### **5.1. Introduction**

This chapter will focus on the community health assessment domain and the types of information technology currently used for numerical-spatial problem solving. Before SOVAT is made available to health assessment researchers, it is imperative that current methodologies involved in the analysis of numerical-spatial problem solving are understood. The focus of this study is the use of information technology for CHA work; that is, “What types of information technology applications are currently being utilized by active researchers for health and population analysis of geographic communities?” This chapter examines the implementation of a survey for identifying information technology use during numerical-spatial problem solving in community health assessment analysis. The identification of existing IT tools will determine what to evaluate against SOVAT for community health assessment problem solving. Before discussion of the survey, community health assessment will be analyzed so that the domain is well understood.

#### **5.1.1. Community Health Assessment**

As defined by Sharma [71], community health assessment “is a process of understanding the communities’ perception of priority health issues in conjunction with the objective collection and analysis of health status data” (p.20). The author goes on to specify that a process-focused approach in conducting a community health assessment entails “Definitions of a community; definitions of health; and development of a model of community health determinants” (p. 20, [71]). The Institute of Medicine has described community health assessment as “all activities

involved in the concept of community diagnosis, such as surveillance, identifying needs, analyzing the causes of problems, collecting and interpreting data, case-finding, monitoring and forecasting trends, research, and evaluating outcomes” (p.21, [71]). Sharma, adapting a definition from Sadan and Churchman [72], goes further and defines community health by a series of dimensions, ranging from process-oriented CHA (previously defined) to product-oriented CHA; however further discussion of this is beyond the scope of this chapter.

### **5.1.2. Initiators of Community Health Assessments**

There are many different organizations that spearhead community health assessment projects. Many of these in the public sector are motivated by national or worldwide benchmark initiatives such as the World Health Organization’s (WHO) *Healthy Cities/Healthy Community* program as well as the United States Department of Health and Human Services (HHS) *Healthy People 2010* initiative. The most widely recognized players in these community health assessments are probably local or state health departments or other state-wide agencies. These organizations are responsible on collecting and reporting on the health of the communities in which they serve. These reports are then made available to the general public.

Another CHA initiative is the result of Community-Oriented Primary Care (COPC) which evolves from healthcare providers such as hospitals and health maintenance organizations (HMOs) [73]. The purpose here is to identify health factors among the local community or provider service area. This CHA initiative is concerned with using local health and population data to understand the origins of local health problems [73]. There are many reasons for why HMOs conduct health assessments. First, as with health departments, they have a responsibility for the health of the populations they serve [74]. Probably more obvious is the financial implications for conducting assessments. With capitation payment, HMOs have an incentive to

keep their members healthy, and thus assessments will provide them with information on what preventive measures are most appropriate for their service area population [74].

The Institute of Medicine's (IOM) 1996 report "Improving Health in the Community" suggested that health assessments would benefit from collaboration between HMOs and health departments [74]. The most suggested area of collaboration pertained to the "design and implementation of data sharing projects" [74]. For example, the Massachusetts Health Assessment Partnership (MHAP) was organized in 1997 between the four largest state HMOs (at the time), three government health agencies, and one health-related nonprofit organization [74]. Since its inception, MHAP has launched six separate assessment projects which are driven by linkage of health and population data sources from the participating institutions [74].

Along the lines of COPC, community health assessment also plays a significant part in public health nursing initiatives. Nursing public health is a recognized field and has been established as such since the mid-1960s [75]. There are at least six defined levels of focus within public health nursing, ranging from assessment and care of an individual outside of an established healthcare setting to analysis and study of community health issues [75]. The latter scope, focusing on the overall health of the community, requires public health nurses to analyze health and population data. The assessments are important for the practice of nursing because they provide nurses with health information about possible patient sub-groups and provide them the opportunity to educate these subgroups using health prevention measures [75]. For example, if a nurse is treating an Italian-American and based on her community health assessment knows that this sub-group of the population has a higher risk of coronary artery disease; the nurse might go out of her way to educate the patient about a healthy diet and lifestyle.

Private foundations are also contributors to the health assessment process. For example, the Birmingham Foundation is a non-profit independent organization dedicated to the health of the residents of Pittsburgh's Southside [76]. Recently the foundation published a CHA report [77]. Fifteen Pittsburgh neighborhoods were included in this 1999 assessment. Some of the measures they reported on included: death rates for stroke and coronary artery disease, death rates for lung and breast cancer, and incidence rates for sexually transmitted diseases (STDs) [77].

Finally, universities and other academic institutions also take on community health assessment initiatives. Outside agencies may commission behavioral and community health professors to conduct assessments of a specific population. Other times, the initiative comes within the university in an effort to identify health issues of the community in which the university or college is located.

### **5.1.3. Components within a Community Health Assessment**

There is no defined standard for how to conduct community health assessments. The Institute of Medicine, in their 1997 report "Improving the Health in the Community" defined a Community Health Improvement Process (CHIP) as consisting of individual cycles for identifying community health issues and implementing community programs [73]. More simply, most CHAs contain the following steps: [78]

1. Identify and define geographic community
2. Identify key players
3. Review and collection of data
4. Establish health assessment priorities for the community
5. Establish a community health plan for the community
6. Implement community health plan for the community

## 7. Evaluate the community health assessment

This chapter will focus exclusively on step 3; review and collection of data. While the other steps are very important, the purpose of SOVAT is to aid in the problem solving of numerical-spatial decision support involving health and population data. As might be evident from these steps, community health assessments may involve many different individuals carrying out many different roles. The representative users of SOVAT are thus the individuals whose roles are to analyze and review the health and population data for the CHA. These individuals might include: individuals within local/state health departments such as biostatisticians and epidemiology data managers; data analysts within business or academic research settings; university faculty/researchers; data analysts in non-profit foundations; and data analysts in healthcare institutions (hospitals, HMOs, etc.).

### 5.2. Background

It has not been determined what types of information technology are used for the review and analysis of health and population data during community health assessments. It is believed that independent applications such as certain spreadsheets, minor database and statistical applications, and even GIS, might be routinely utilized. A few publications described the use of GIS during health and population analysis [79-81], however GIS seems to be utilized more in environmental health analysis. There have been no measurements such as published surveys that have focused on the use of information technology during analysis of health and population data. The closest survey found was done by a group in Canada in relation to information technology needs assessment and was discussed in [82]. This survey was sent to 30 community health professionals throughout Canada and was meant to gauge their interest and need for OLAP and

GIS technology during community health assessment. The survey was not designed to identify specific information technology used during numerical-spatial problem solving.

Another relevant survey was developed as a result of the Turning Point program, which is a joint initiative of the Robert Wood Johnson Foundation and the W.K. Kellogg foundation to “transform and strengthen the public health of the United States by making it more community based and collaborative” [83]. One of their five collaboratives is the Information Technology Collaborative (ITC) whose mission is to “assess, evaluate, and recommend to national policy-makers innovative ways to improve the nation's public health infrastructure by utilizing information technology to effectively collect, analyze, and disseminate information; by improving data access and community participation for making public health decisions; and by enhancing the performance of the public health system through the use of information technology” (webpage, [84]). In 2001, the ITC commissioned a survey of local health departments to “inventory and evaluate their current use of IT and their perceived IT needs” (p. 124, [85]). The survey was designed to answer 3 main questions:

1. “What information technology is being used in local US health departments?
  2. How do end-users of the technology, meaning professional staff members in local health departments, rate the software they use?
  3. What are the information technology needs of local health department staff members?”
- (p. 125, [85])

The survey was intended for four types of health department professionals: an administrative person, medical/clinical person, an IT person, and an environmental/sanitation person. Three thousand one hundred thirty one surveys were sent through the mail in 2002, with 344 representing the final usable number that was returned (for a response rate of 11%). More than

500 different types of software programs were identified from the survey [85]. The single most software mentioned was Microsoft Access, with Microsoft Word and Microsoft Excel representing the second and third most popular software. For environmental health staff, Access was the most popular. The survey also broke information technology down by 10 essential services in public health: monitoring health status, diagnosing and investigating, informing/educating/empowering, mobilizing community partnerships, developing policies and plans, enforcing laws and regulations, linking people to needed health services, competent workforce, evaluating services, and research. Under the category of monitoring health status, KIPHS (Kansas Integrated Public Health System) was listed as the most popular software, followed by Excel, WIR (Wisconsin Immunization Registry), SPSS, and Access. Under the service of “diagnosis and investigate” the most popular was Access followed by Epi-Info (a statistical analysis tool), Healthspace, Word and Excel (tie for 5<sup>th</sup> place).

Another survey was established by The Washington State Department of Health. Within the department is a Public Health Improvement Partnership (PHIP) Information Technology Committee. The committee focuses on information needs for public health professionals and policy makers [86]. One of the tasks undertaken by the committee was to conduct a technology inventory via a survey to better understand the applications used by public health professionals across the state [86]. The survey, which is published on their website [87], does ask health officials to identify information technology used, however there is no mention of numerical-spatial problem solving tasks. The survey asks them to “list up to 10 Epidemiology , Surveillance, and Assessment custom software applications that are most useful to your LHJ [Local Health Jurisdiction] and that could be perhaps useful to another LHJ” (webpage, [87]).

The survey then asks the respondent to answer questions for each specific software application listed. This request is very general in that it does not identify the IT used during the *process* of numerical-spatial problem solving. There is nothing that breaks down the steps within a numerical-spatial problem and gathers information on the IT used during these scenarios. The survey results are not publicly disclosed as per the wishes of the Washington state department of health [88].

Other public-interest groups are conducting similar surveys as well. The National Association for County and City Public Health Officials (NACCHO) and the Centers for Disease Control (CDC) are implementing a survey to be sent to Local Public Health Agencies (LPHAs) [89]. The 3 previous studies were done in 1997, 1993, and 1990 [89]. A portion of the survey is related to community health assessment and planning. One of the questions within this section relates to the use of certain community health assessment planning tools such as, MAPP<sup>20</sup>, APEX PH<sup>21</sup>, PACE EH<sup>22</sup>, NPHPSP<sup>23</sup>, and PATCH<sup>24</sup>. These resources help individuals plan and organize their assessment. None of these ‘tools’ (these are mostly reports) focus on information technology use during analysis of health and population data.

The NPHPSP, the National Public Health Performance Standards Program is a collaborative effort to develop performance standards for state and local public health systems [90]. Part of the initiative includes a survey sent to both local and state agencies in order to gauge their performance to the established standards. One of the topics (performance indicators) on the local public health questionnaires is “Access to Utilization of Current Technology”. This indicator

---

<sup>20</sup> Developed by NACCHO and the Centers for Disease Control (CDC, [http://mapp.naccho.org/mapp\\_introduction.asp](http://mapp.naccho.org/mapp_introduction.asp)) 06/30/05

<sup>21</sup> Developed by NACCHO (<http://www.naccho.org/topics/infrastructure/APEXPH.cfm>) 07/01/05.

<sup>22</sup> Developed by NACCHO (<http://www.naccho.org/topics/environmental/CEHA.cfm>) 07/01/05.

<sup>23</sup> Developed by the CDC (<http://www.phppo.cdc.gov/nphpsp/>) 07/01/05.

<sup>24</sup> Developed by the CDC (<http://www.cdc.gov/nccdphp/patch/>) 06/30/05

contains a question, “Does the LPHS [Local Public Health System] use GIS”? Unfortunately only indicator-level results are available, thus special permission is needed for the results of this specific question [91]. There is however some public information data aggregated by indicator. The published results show that the local public health agencies received a “30.88” compliance score (out of 100) for the indicator “Access to and Utilization of Current Technology” [92]. This score was the second lowest out of the 31 total (sub) indicator scores, suggesting significant deficiency in IT utilization among local public health agencies. The data is from 2002-2005 and represents information collected from 315 local public health systems located across the country [93].

### **5.3. Methodology**

While the previous surveys are valuable at identifying information technology utilization at the health department level, they do not specifically address the practice of community health assessment and numerical-spatial problem solving. As mentioned earlier, community health assessments are not only performed at health departments. Thus, in order to identify information technology for numerical-spatial problem solving in community health assessments, a survey must be designed for any individual at any type of institution (and not just health departments). In order to truly uncover IT utilization, a new survey needed to be developed. The survey is shown in **Appendix J**.

The survey was electronic (accessed through the Web). The survey itself contains 5 parts:

1. Background section: gathers background information on the participant such as whether they have completed a health assessment in the last 3 years, and what type of

organization they work at (choices included: health department, healthcare institution, university, private foundation, non-governmental organization, and other).

2. Scenario section: describes the imaginary scenario that the participants should think about when completing the survey. That is, they are given electronic data sets and are asked to analyze the health and population data by completing 5 community health assessment tasks.
3. Survey Section - Part 1: describes the 5 tasks that they would need to solve. Below each task is a text box. The participants are asked to describe how they would complete the task given the electronic data available to them. They are urged to think about what information technology applications (software, web-based tools, etc.) they would use to solve the task.
4. Survey Section - Part 2: describes the same 5 tasks from part 1. This time, the tasks are broken out into steps (validated by a community health expert (RKS)). Next to each step is a small text box, where the participants are asked to think about what specific IT application they would use for the particular step.
5. Survey Section - Part 3: asks the participants to note any additional IT that they did not list in the first two parts. They were given a text box to write their responses.

For assessing content validity, the survey was sent to three experts in community health assessment. They were asked to review the survey and assess whether the tasks and the nature of the survey (especially the scenario section) was appropriate for community health professionals. Overall, the experts felt the content within the survey was appropriate. Their feedback was used to create the final version of the survey.

### **5.3.1. Data Collection**

The survey was web-based and thus the participants were contacted via an email cover letter (**Appendix K**). The cover letter described the survey; the types of participants encouraged to complete the survey, the compensation amount, and the URL to access the survey. The cover letter was similar to the survey introduction page, however was not as detailed in describing the survey. It was decided to send the survey to anyone at any organization that might have conducted a community health assessment, commissioned a community health assessment, or used community health data to develop their own community-based program. The preferred time frame for these activities was within the last 3 years. This time frame is identical to the one used in the NACCHO survey described earlier [89]. The survey was estimated to take approximately 20 minutes to complete (based on the completion time of one of the community health experts who analyzed the survey for content validity). For their time, participants were sent \$5.

Convenience sampling as described in [94] was used for recruitment. The researcher responsible for sending out the survey (MS) searched the Web, the literature, and utilized contact names for anyone who seemed to be associated with the process of analyzing health and population data for community health assessments. Among the types of participants who were emailed were: health department employees such as biostatisticians/epidemiologists, university researchers, industry consultants, and community health analysts within governmental organizations. Each contact was sent the email cover letter with the subject line: “A Survey to Identify Information Technology (IT) in Community Health Assessment Research”. Follow up was done 3 weeks later.

## 5.4. Results

Around 500 emails were sent out and 27 responses were received (~ 5% response rate). This low rate was anticipated since most of these contacts were identified through literature and Web searches. The recruitment goal prior to the study was around 30 (+/- 5). This matches the number of participants in the Canadian survey which is the most similar to this survey [82]. Braithwaite conducted a literature of online surveys to health professionals and found the response rate range to be between 9 – 94% for 12 surveys between 1999 and 2002 [95]. It is believed that the estimated length of completion (20 minutes) and the requirement for free text entry related to complex numerical-spatial tasks might have contributed to this lower rate. Also, the fact that recruitment was via an email introduced from an unknown entity brings the potential of response-related problems such as the higher probability of the email being discarded by a spam filter or otherwise ignored by the recipient amongst a large pile of other inbox message items.

Of the individuals deciding not to respond, some of the reasons given were: “The survey is not applicable to me”, “This is survey is too technical”, “I do not have the time to complete this”, and “I will forward this to my colleague who will be a better person to complete it”. These responses suggest that the ones who responded are the right group and that the low response is consistent with the goal of the study to focus on numerical-spatial problem solving.

All the participants who completed the survey either had performed at least one of the following activities within the last 3 years: Conducted their own community health assessment, used data from a health assessment to develop their own initiative, or commissioned a community health assessment. A valid response was any survey that was completed by someone involved with community health assessment data in the last 3 years and thus would be able to

answer “Yes” to at least one of the three background questions. Table 5-1 shows the breakdown by organization.

**Table 5-1: Twenty seven survey participants by type of organization.**

<b>Organization</b>	<b>Number of Survey Participants</b>
Health Department (local or state)	14
Other (Government agencies)	6
University	4
Healthcare Institution (Hospital, HMO, etc.)	2
Private Institution	1
Non-Government Organization	0

The completed surveys were analyzed using survey section Part 1 and Part 2. Part 2 was very straightforward since the responses here were one or two-word answers. Part 1, the free-text portion, was read and analyzed to confirm or add to the responses given from Part 2. In Part 2, a specific software name (or names if more than one) was typically used at each step. The specific software was recorded, but for analysis purposes, the types of software were aggregated into IT groups. For examples, SAS, SPSS, Stata, and Epi-Info were grouped into the category of “statistical software”. ArcGIS and Epi-Map were grouped into the category of “GIS software”. Web resources such as online vital statistic data sets, and data analysis tools (that perform calculations) were grouped into the category “Web-based Interface”. For example Figure 5-1 shows an example of a response in Part 2 of the survey. Here, the response for step 1 is “use the state website to run queries”. This response was recorded, and then grouped into the higher category “Web-based Interface”. The response for the third step is “Arc GIS 9.1”. This response was recorded, and then grouped into the higher category “GIS Software”. The final step would add one to the grouping “Statistical Software” (for the response SAS 9.1.3) and one to the grouping “GIS Software” (for the response ArcGIS 9.1).

## Part 2

---

### Task 1

#### Step

*Access County level data*

*Find deaths/100,000 in 1996*

*Identify bordering counties*

*Compare Border Counties to Allegheny County*

#### Tool(s) Used

Use the state website to run queries.

Same as above

Arc GIS 9.1

SAS9.1.3 and ArcGIS9.1

**Figure 5-1: Example responses from Part 2 of the survey.**

The reason for the grouping was to reflect the purpose of the survey; to identify the types of information technology and not necessarily the specific individual applications. It was felt that analyzing the results by individual application put too much emphasis on the use of specific software rather than the type of software. For example, an individual might use Stata not because he/she prefers Stata over SAS, but because this is the application provided by his/her organization or this is the application in which he/she receives the biggest financial discount. The use of the brand is not as important as the use of the type of software. Survey participants were asked to be as specific as possible (including version number) in order to facilitate grouping of the individual applications into their higher category.

Once all the responses were analyzed, the data were summed by task. Each step then received one IT group that was the highest for that step. For example, in Part 2 Task 1 - Step1, the totals were calculated and the most frequent IT used in that step was “statistical software” with 11 responses, followed closely by “web-based interface” with 8. It was decided to group the steps into categories by type of numerical-spatial problem solving step. The groups were determined to be: Data Management/Access, Data Navigation, Geographic Comparison, Spatial

Boundaries, Spatial Modeling, and Ranking Analysis. The grouping of each step into these categories is shown in **Appendix L**. The most popular IT overall within the step category was simply the IT that occurred most frequently in the group (ex. Under the Geographic Comparison category, “statistical software” is the most popular tool in all 5 of the steps in this category, so it is considered the most popular overall for “Geographic Comparison”). These results are shown in Table 5-2.

**Table 5-2: Problem solving category and the most popular IT for that category.**

<b>Numerical-Spatial Problem Solving Category</b>	<b>Most Popular Type of IT</b>
Data Management/Access	Statistical Software
Data Navigation	Statistical Software
Geographic Comparison	Statistical Software
Spatial Boundaries	GIS Software
Spatial Modeling	Statistical Software
Ranking Analysis	Statistical Software

Clearly, the data indicates that statistical software is the most popular technology for most types of (numerical) problem-solving purposes. GIS is the most popular for spatial boundaries (i.e. does one county border another county). This is consistent with the literature which suggests that while gaining in popularity, GIS is not used to its full potential as researchers use it primarily for only spatial display and simple functions [82, 96] and not for more complex spatial analysis during numerical-spatial problem solving.

Web-based interface tools were popular for data management/access; however, it was evident from analyzing survey section Part 1 that these tools were used mostly for accessing the data while the management portion of the process would be to use a statistical software package. Thus, the responses from survey section Part 1 indicated that they used a web-based feature to view and download the data, and then used a statistical software package for the data management aspect. The focus of the scenario was not on data access (i.e., in the scenario

described on the survey, the data is already available to them in electronic form) but rather on the management of the data.

Since statistical software and GIS software constituted the most popular types across these categories, it was decided to examine the breakdown by specific application. This is why all individual software was recorded before being aggregated in software type. Tables 5-3 and 5-4 show the breakdown by these two different types of software. Included with statistical software is Excel since it is commonly used in place of statistical software for numerical problem solving. The conventional map (considered as either a paper map or a map on a Web site such as Yahoo Maps) is included with the GIS software because it is a common substitute.

**Table 5-3: Technology for numerical problem solving steps.**

<b>Number of Instances of Statistical Software Technologies</b>	
SPSS	11
SAS	6
Stata	3
Epi-Info	3
Excel	13

Note: A participant indicating they use both SAS and SPSS, for example, would be scored “1” for SAS” and “1” for SPSS”.

**Table 5-4: Technology for spatial problem solving steps.**

<b>Number of Instances of GIS Software Technologies</b>	
ArcGIS	12
Epi-Map	1
Forestry GIS (fGis)	1
Conventional Map	8

Note: A participant indicating they use both ArcGIS and Epi-Map, for example, would be scored “1” for ArcGIS and “1” for “Epi- Map”

Table 5-3 suggests that statistical software is used more than Excel (even though Excel is more popular than any tool by itself). For example, 23 out of 27 participants use statistical software (85%) where 13 out of 27 used Excel (48%) for community health assessment problem

solving. Examining the individual packages, SPSS was the most popular statistical package (with 11), followed by SAS (with 6). Examining the GIS table (Table 5-4), GIS packages were used more often than a conventional map. In fact, 14 of the 27 participants used GIS software (52%) where as only 8 out of 27 used a conventional map (30%). The most popular application was ArcGIS (with 12).

Table 5-5 shows the participants who use statistical software.

**Table 5-5: IT applications used with statistical software for community health assessment.**

	<b>Participants Who used Statistical Software with other Tools (N=18)</b>
GIS	12
Web-based Interface	8
Excel	7

*Note: A participant, who indicated they used Statistical software with GIS, and Excel, would count as “1” for GIS, and “1” for Excel.*

**Table 5-6: IT applications used with GIS software use for community health assessment.**

	<b>Participants Who used GIS with other Tools (N=14)</b>
Statistical Software	12
Web-based Interface	6
Excel	5

This same type of analysis can be applied to participants who use GIS (N=14). This is shown in Table 5-6

In order to get a sense of the different types of users of information technology in community health, the Apriori algorithm was run which generated association rules among the individual software application (**Appendix M**). The Apriori algorithm is a well know algorithm developed by Agrawal and Srikant for mining association rules within data sets [97]. The rules demonstrate which individual applications are used together. SPSS and ArcGIS show the

strongest link, in that someone who uses SPSS will most likely use ArcGIS to solve numerical-spatial problems. SPSS is a relatively simple statistical software package and contains a lot of visual presentation; spreadsheet, tables, bar graphs, plots, etc. ArcGIS is used for visual presentation via spatial display. Thus someone who uses SPSS and ArcGIS seemingly prefers software that provides useful visual environments, rather than complex command-line features. SAS has the strongest link with Web-based interface and the use of a paper map. This represents a more advanced user. This relationship indicates that the data is downloaded from the Web and then the majority of the work is done with SAS, without the reliance on visual display. The user feels most comfortable in the command-line environment working with the data, and only uses a paper map when analyzing spatial boundaries (which is something SAS cannot do). Everything else is done in the SAS command-line environment.

## **5.5. Discussion**

The previous data suggests a strong relationship between GIS and statistical software. Thus if a participant is using more than one type of information technology for community health assessment analysis, it is most likely a combination of GIS and statistical software. Examining the responses specifically of the participants who use both GIS and statistical software highlights the relationship between these two technologies during numerical-spatial problem solving.

*Part 1*

*Task 1*

*"How does the deaths/100,000 of Allegheny County in 1996 compare to the deaths/100,000 of each of the counties that border it?"*

1. I would use ArcGIS (v9.1) to open a state map and identify the counties that border Allegheny county.
2. Next, using SPSS v13.0, I would aggregate the death file and population file to get annual death totals and populations for counties by linking the geography table to the death and population tables.
3. Next I would compute the age-adjusted death rates for counties by year using the tables generated in the previous step (SPSS v13.0).
4. Finally I would generate a report using SPSS v13.0 of 1996 death rates by county after selecting Allegheny and its bordering counties.

**Figure 5-2: CHA professional who uses both GIS and statistical software.**

Figures 5-2 and 5-3 are survey responses from a participant who uses statistical software and GIS for numerical-spatial problem solving. Here, the participant's responses for task 1 in both Part 1 and Part 2 of the survey are shown. Reviewing both parts provides an excellent description of how GIS and statistical software are used to solve numerical-spatial problems.

*Part 2*

*Task 1*

<u>Step</u>	<u>Tool(s) Used</u>
<i>Access County level data</i>	SPSS v 13.0
<i>Find deaths/100,000 in 1996</i>	SPSS v 13.0
<i>Identify bordering counties</i>	ArcGIS 9.1
<i>Compare Border Counties to Allegheny County</i>	ArcGIS 9.1

**Figure 5-3: Same participant and the response for task 1 for Part 2 of the survey.**

Data is loaded into SPSS for analysis. Data navigation and aggregation is done using SPSS to determine the deaths/100,000. GIS is then used for spatial display to determine the counties that border Allegheny County. Since the analysis has been done with SPSS, the participant

seemingly then transfers the data into ArcView and uses spatial display to analyze the rates for Allegheny County versus its bordering areas.

This response shows how statistical software and GIS are used for numerical-spatial problem solving. Statistical software is frequently used for analysis and navigation-related problem solving steps, while GIS is used for simple spatial display. As researchers begin to feel more comfortable with GIS technology, most likely GIS will constitute a more important component within community health assessment analysis. Health departments and universities are purchasing GIS software packages because they realize their potential. However, end users are not utilizing their full potential; preferring to use statistical software to do the brunt of the work, and using GIS for display and reporting purposes.

#### **5.5.1. Public Health Curriculum**

The considerable use of statistical software is not surprising. There is a significant reliance on these tools for analysis in Epidemiology and public health-related research. Public health curriculums including courses in biostatistics typically require students to utilize some sort of statistical software application for assignments. The students then learn by doing and become comfortable using these types of software. GIS, on the other hand, is not nearly as popular in public health curriculums. Only a handful of schools teach GIS courses in relation to public health. If a university does offer a GIS course, it is most likely through a geology or information and computer science department. Public health students might not feel compelled or even comfortable taking these courses outside of their domain. Public health curriculums shape the future community health assessment professionals. In order for GIS to gain in popularity, it is important for these programs to implement GIS into their programs and teach students how to use them in a hands-on manner as is done with statistical software.

## **5.6. Limitations**

The main limitation with the survey is the omission of reliability and validity analysis. This was due mainly to time constraints of the study. Content validity was used, as a draft of the survey was distributed to 3 community health assessment experts. Beyond that, there was no validation of construct validity (i.e., the survey is measuring what it is intended to measure) and reliability (i.e., the survey is consistent in measurement). It was felt however, that the survey was well-developed and was able to appropriately record information technology use in numerical-spatial problem solving.

## **5.7. Conclusion**

This chapter described the effort to determine the utilization of information technology during community health assessment analysis. A survey was constructed that asked participants to describe the tools they use to solve typical numerical-spatial problems during community health assessment research. The survey was electronic (on the Web) and sent to approximately 500 individuals who seemed to be associated with community health assessments in some fashion. The nature of the survey (time, free-text entry, complex tasks) produced a low response rate (~5%) as 27 surveys were analyzed. The number was within the pre-study goal of 30 +/- 5. Participants included individuals at local and state health departments, universities, government agencies, and healthcare organizations. Survey background data indicated that within the last three years, all the participants had either: conducted a community health assessment, commissioned a community health assessment, or utilized data from a community health assessment for their policy development and planning. Responses indicated a frequent use of

statistical software applications such as SPSS and SAS. Excel was the most popular single information technology component. However, when combining individual applications into groups, “statistical software” was used more often than Excel for data navigation, data management, geographic comparison, spatial modeling, and ranking analysis purposes. The most popular statistical software package was SPSS. GIS was the most popular for purposes of spatial boundary detection. This supports prior research efforts that conclude that GIS is underutilized for research purposes and is only used for simple spatial display of the data. The most popular GIS software was ArcView.

The relationship between statistical software and GIS for numerical-spatial problem solving shows that the driving force behind numerical-spatial analysis is statistical software. Thus, data management, data navigation, and geographic comparison are done with SAS, SPSS, or Stata and then GIS is used to display a digital map of the corresponding area. For this purpose, the data may or may not be imported into GIS from a statistical package since the only function is for spatial display.

The purpose of this survey was to identify the information technology used for numerical-spatial problem solving. The goal was to take the most popular technologies and use them as the “control” on which to evaluate against SOVAT during the next study; the summative evaluation of SOVAT. As mentioned, while head-to-head Excel was the most popular, statistical software (after grouping individual technologies) was used by more participants than Excel. In addition when analyzing by type of problem solving step, statistical software was the most popular for every category except spatial boundary detection (where GIS was most popular). The conclusion is to use statistical software and GIS as the technology on which to evaluate against SOVAT. Specifically, the most popular tools within these two categories, SPSS and ArcView, will be

used. For the remainder of this dissertation, the combined use of these technologies will be referred to as “SPSS-GIS”. It is believed that this combination best represents the current nature of numerical-spatial problem solving for community health assessment research.

## **6. EVALUATION OF SOVAT FOR NUMERICAL-SPATIAL PROBLEM SOLVING IN COMMUNITY HEALTH ASSESSMENT RESEARCH**

### **6.1. Introduction**

The usability evaluation of SOVAT detailed in chapter 4 was essential in identifying issues with human-computer interaction. The end result of this rigorous process is a decision support system that is conceivably more usable than it was before the study began. The usability study was an essential step in the process of analyzing an OLAP-GIS system for numerical-spatial problem solving in community health. It was not, however, the final step. Much emphasis in this document has been on the uniqueness of combining OLAP and GIS. This uniqueness can not be overlooked or underestimated. The characteristics and properties of these two underlying technologies create a very different type of decision support system. On-Line Analytical Processing (OLAP) adds an entire new dimension to navigation, data exploration, and data management that traditional databases do not support. As discussed, it is believed that these capabilities will provide a more powerful, useful, and valuable system to community health researchers. As science mandates, these hypotheses must be supported (or not supported) by valid research evaluation. This entails moving from a needs-driven analysis (as with the usability study) to a hypothesis-driven study (summative evaluation). This is the final proposed step in relation to this work. The usability study in chapter 4 has presumably brought the interface up to an acceptable level in relation to human-computer interaction. The purpose of the survey described in chapter 5 was to determine the existing information technology tools that are used for numerical-spatial problem solving in community health assessment analysis. These tools will serve as the “control system” for this final summative evaluation. Evaluating SOVAT against these existing technologies will provide some perspective on how an OLAP-GIS

compares to the current methods of numerical-spatial problem solving in community health. If the objective and subjective data suggests that SOVAT is superior to the “control system”, it will provide a positive outlook on the impact of an OLAP-GIS system for numerical-spatial problem solving.

## **6.2. Background**

Chapter 2 detailed the uniqueness of an OLAP-GIS decision support system for numerical-spatial problem solving. Because of this uniqueness, there has been no prior study that has evaluated an OLAP-GIS system for numerical-spatial problem solving. In fact, little research has been done in the emerging field of public health informatics in relation to evaluation of systems for numerical-spatial problem solving. The closest study is an unpublished work by Edward Bunker at Johns Hopkins University [98]. Bunker created a visualization tool for analyzing public health data. The emphasis of the system is on different methods of visualization for public health analysis. Much less focus is on database components. In developing his visualization tool, Bunker aimed to create an interface that supported the analysis of public health data through the notion of “person, place, and time” [98]. His interface contains three separate visualization techniques for these three notions. For example, to demonstrate demographic information of an individual (person), Bunker uses a population pyramid. For place, the interface contains a map, and for time, the interface displays a line graph. Bunker’s research hypothesis related to the use of his visualization tool by public health practitioners. More specifically, he aimed to determine if his prototype was better than a standard spreadsheet (such as Excel). Bunker chose the spreadsheet because he felt that it was the most popular tool used by public health practitioners to analyze public health data (*Note: This is based on personal*

*communication with the researcher, Bethesda, MD, July , 2005.*). No empirical data such as a survey was used to support this claim. A within subjects crossover study design was implemented. Half of the participants used the spreadsheet first to complete assigned tasks and then switched and used the visualization prototype. The other half used the prototype first and used the spreadsheet second. Brief training was given before using both systems. Two tasks were then given (an easy task and a difficult task). Both objective and subjective measures were used for the evaluation. Objective measures included accuracy and time to complete the tasks, while satisfaction was used as a subjective measure. Time and accuracy were measured as participants completed assigned tasks. After completing the tasks, the participants completed the IBM PSSUQ questionnaire which, as described, measures user satisfaction. The participation criteria included current or recent public health practitioners who had at least two years experience in public health working with public health data. Participants were from the Baltimore/Washington, D.C. area. In total 24 participants were enrolled in the study (15 males and 9 females). Thus 12 used the spreadsheet first, while the other 12 used the prototype. Then, for the second session, the groups switched software.

The results from the study all favored the spreadsheet over the researcher's prototype. Accuracy was 10% lower for the prototype than the spreadsheet [98]. Time was about 50% longer with the prototype than the spreadsheet. The subjective measure using the IBM PSSUQ showed that satisfaction was higher with the spreadsheet. Thus both the objective and subjective data indicate that the spreadsheet was better than the researcher's visualization prototype.

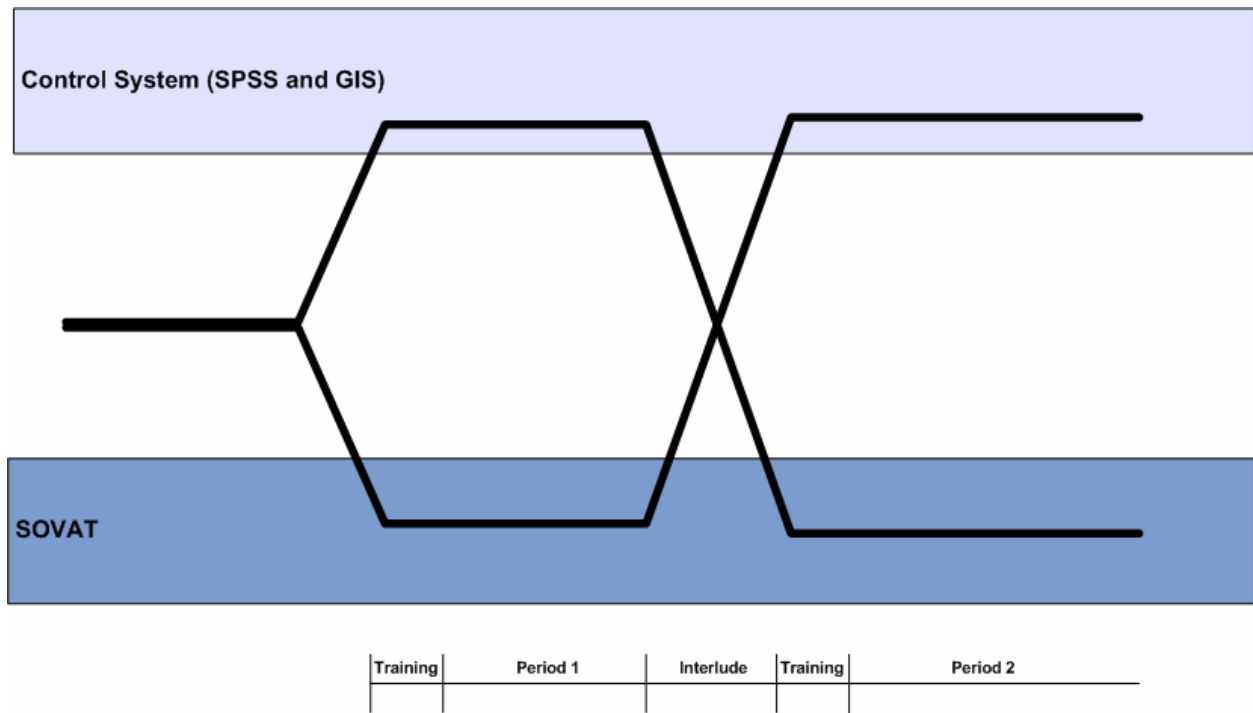
While not in the public health domain, another study worth noting by Hart [99] analyzed On-Line Analytical Processing in relation to perceived usefulness and Davis's Technology Acceptance Model (TAM) [100]. The study is important because it takes OLAP out of its

traditional complex and technical focus and discusses its relation to a cognitive and user-centered angle. This is the primary angle in which SOVAT will be evaluated. In addition, few studies have focused on technology adoption in relation to business intelligence. The authors phoned information technology managers of South African companies and determined if they used OLAP in their daily work. Those that did were sent a questionnaire that was based off of Davis's TAM questionnaire [69]. Fifty six total participants were enrolled in the study. Ninety percent of the participants had used OLAP for at least 2 years. The results from the questionnaire indicated that perceived ease of use (as well as other constructs) was positively correlated with their perceived usefulness of OLAP.

The rest of this chapter will focus on the design of the summative evaluation of SOVAT.

### **6.3. Methodology**

The design of this evaluation used a very similar approach to the one implemented by Bunker [98] as well as Zeng [101]. It was a within-subjects crossover study. Participants used both SOVAT and the combination of SPSS and GIS (referred to here as the "control system" or "SPSS-GIS"). Intervention was randomly assigned. This design is illustrated in Figure 6-1 (adapted from [101], pg. 109).



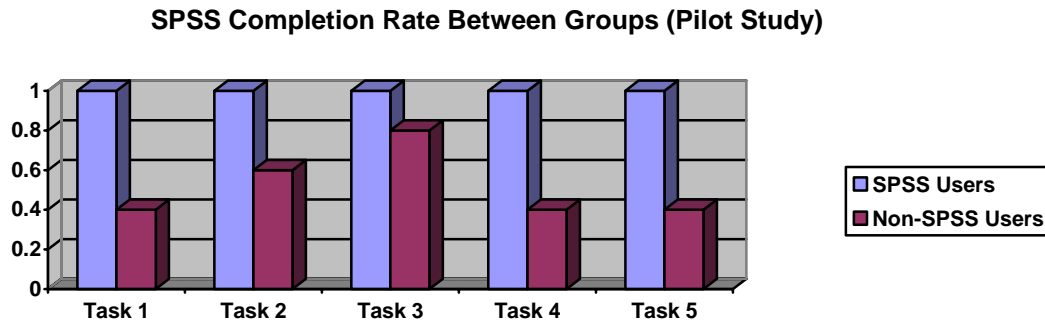
**Figure 6-1: Research Design of the SOVAT evaluation study**

Depending on which order the participants were assigned, they either used SOVAT during period 1, or SPSS-GIS, given an interlude period between 2 – 3 weeks, and then used the other system. The training occurred before the start of the study (before data was recorded). Training was an important aspect of this design. Since two systems are being evaluated (a “control” system and a new system) it was felt that establishing a certain level of proficiency for both systems was necessary for eliminating confounding variables related to system familiarity. For example, SOVAT is new, and thus no participant had experience using it. Thus, it was essential that training is given before the participants were asked to evaluate it. Additionally, training was essential in getting the participants up to speed on using SPSS for solving numerical-spatial problems. Using SPSS for this type of problem solving is much different than the manner in which most individuals use SPSS when performing descriptive and inferential statistical analysis. Certain functions required during numerical-spatial problem solving would likely be new to

them. Conceivably, participants receiving training on both systems before evaluation would eliminate any factors related to proficiency.

### **6.3.1. Pilot Study**

A pilot study was implemented to determine the appropriate users for the evaluation. Due to the fact that SPSS is very error prone and requires a certain level of understanding of its functions and layout, it was decided to implement a pilot study in order to verify that proficient SPSS users should be utilized for the evaluation. Three SPSS users and five non-SPSS users (with little or no experience) were recruited (it was desired to have 4 and 4, however one participant claimed to have experience with SPSS prior to the study and it was evident from the tasks that this was not the case). The pilot study design was the same as the summative evaluation as outlined in Figure 6-1. It was determined that the most appropriate participants were experienced SPSS users. This was because many of the non-SPSS users had difficulty completing the tasks and became extremely frustrated with the study. Figure 6-2 shows the SPSS completion rate between the groups for the five tasks. The experienced SPSS users (N=3) were able to complete all of the tasks using SPSS, while at least one of the non-SPSS users (N=5) failed for each of the tasks. In order to properly evaluate SOVAT, it was decided to use participants with experience using SPSS.



**Figure 6-2: Rate of completion (between groups) using SPSS during the pilot study.**

### **6.3.2. Recruitment and Setting**

Thirteen individuals participated in the study. The participants included graduate students, professors, and some non-student researchers within the health science schools at the University of Pittsburgh. The specific schools within the University's Health Sciences include the: School of Dental Medicine, School of Medicine, School of Nursing, School of Pharmacy, Graduate School of Public Health, and the School of Health and Rehabilitation Sciences. The essential requirement was that the participants were experienced with using SPSS. Interested participants replying that they had never heard of SPSS or had used it a couple of times, were not enrolled in the study. The graduate students were recruited via fliers that were posted around their school (**Appendix P**).

### **6.3.3. Software used in the Study**

This evaluation involved the use of two systems; the current technology used for numerical-spatial problem solving, and the new system being evaluated. The current technology (SPSS-GIS) comprises two separate software applications: SPSS statistical software 13.0 and ArcView 9.1. During the "control session", participants had access to both SPSS and ArcView to solve the tasks (i.e. both of these two applications were on the computer and available during this session). Every task required them to use both of these applications. The other system being

evaluated was the Spatial OLAP Visualization and Analysis Tool (SOVAT). During the “SOVAT session”, participants had access to SOVAT (and no other application) to solve the tasks (i.e. SOVAT was running on the computer during this session).

#### **6.3.4. Study Procedures**

Before entering the conference room, participants were asked to complete the informed consent form (**Appendix N**). The study lasted approximately 2.5 hours as was divided into two parts: training and then evaluation. Once in the conference room, the participants were shown a pre-recorded instructional video that served as the introductory script for using the system (depending on the system they were using for the current session). They were allowed to take notes during this time. The content of the video, including the facets of the interface and the methodologies for producing queries, was deemed appropriate for use in the study by one of the co-investigators (VM) who is an expert in Human-Computer Interaction (HCI). After watching the video, the participants were given two practice tasks to solve using the system. After completing each task, they were shown a video solution for the particular practice task.

Once the two practice tasks were completed, the participants were asked if they felt that they were able to use the system to complete 5 additional tasks similar in nature to the practice tasks they had just completed. This was an important part of the training content within each session. Participants who answered “yes” were considered to be proficient with the system and allowed to continue to the evaluation. Participants who answered “no” were not considered to be proficient with the system and not allowed to continue to the evaluation. These participants were compensated for their participation and not included in the evaluation of SOVAT. In order to perform a proper evaluation in a crossover design, the participant must have evaluated both SOVAT and the “control” system. That is, they must have answered “yes” to both systems

(indicating that they were proficient in both systems under evaluation). A participant who successfully completed the first period of the crossover, but then was not able to complete the second period (i.e. answers “No” to the question) was removed from the results of the evaluation.

If the participant answered “Yes” and thus deemed proficient with the system, the evaluation then began (As it turned out, all enrolled participants said “yes” and completed both sessions). The participants were then given five problem solving tasks to answer using either SPSS-GIS or SOVAT (**Appendix O**). The tasks represented realistic community health assessment problems, and were deemed appropriate by a co-investigator (RKS) who is an active community health researcher. They consisted of performing local and state-wide comparison of geographic areas, ranking of diseases or geographic areas based on health measures, and defining and comparison of customized geographic communities. For the two systems, it was decided to make the task similar but not identical. So that the participants would not all receive the same ordering of tasks, Balanced Latin Squares (BLS) was used. Participants were randomly assigned to an ordered row of tasks. Camtasia screen capture software was used to record their interaction while the external microphone captured their verbal thoughts. Once the participants completed the 5 tasks, they were asked to complete the IBM PSSUQ satisfaction questionnaire (**Appendix F**). If this was their second session, they were also asked to complete the computer background questionnaire (**Appendix D**).

#### **6.3.5. Objective Measurements**

Two variables were used as objective measurements for this evaluation:

- Time to complete each task – This measure was defined by the time between when a participant finished reading the question to when the participant indicated he/she was

done. The use of screen capture software allows one to measure the participant's time for each task. This screen capture method is also non-intrusive.

- **Answer to Problem** – An answer was defined as the action of the participant verbalizing an answer to all the questions in the task followed by saying that they were 'done'. The answer did not have to be the same as what was currently being shown on the screen at the time. The participant had to answer all parts of the question correctly to successfully answer the task.

### **6.3.6. Subjective Measurements**

As mentioned, the IBM Post-Study System Usability Questionnaire (PSSUQ) (**Appendix F**) was used for subjective measurement. This is the same questionnaire that was used for the usability study as described in section 4.4.4.

A final subjective variable was user preference. This was obtained by conducting a brief post-study interview immediately following the completion of the second session. The question posted to every participant was "Which software system did you like better and why"? User preference was identified from their response.

### **6.3.7. Statistical Analysis**

Both descriptive and inferential statistics were calculated for analysis purposes. Descriptive statistics were used for time, answer, satisfaction, and user preference. Inferential statistics was performed by conducting mixed model analysis. This method enabled for design, period, and intervention effects to be identified across the variables 'time' and 'user satisfaction'. Statistical analysis was conducted using SPSS 13.0 for Windows. The alpha level for this study is .01, and the beta is .80.

## 6.4. Results

### 6.4.1. Objective Measurements

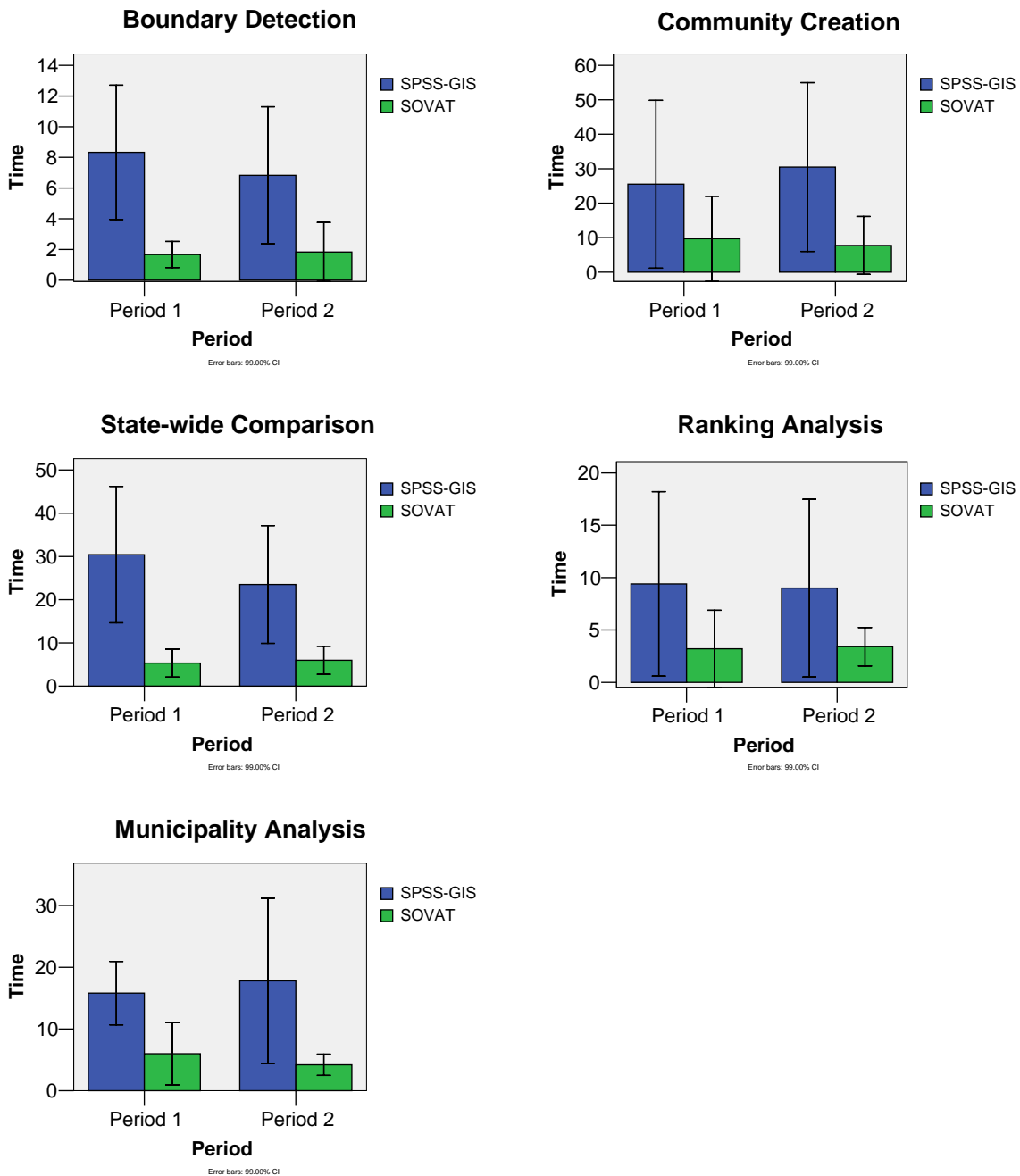
#### 6.4.1.1. Time

Table 6-1 shows the mean and 99% confidence interval for time rounded to the nearest minute. The results are shown by task by period. The 5 tasks are named based on their most distinguishable characteristic and are: boundary detection, community creation, state-wide comparison, ranking analysis, and municipality-level analysis. If a participant did not complete a task using SPSS-GIS, then the corresponding task in SOVAT was not factored into the analysis. For example, if using SPSS-GIS, the participant did not complete the Community Creation task, their time for Community Creation using SOVAT was not factored into the analysis.

**Table 6-1: Mean time (rounded to the nearest minute) and 99% CI per period per task.**

<b>Task</b>		<b>SOVAT</b>	<b>SPSS-GIS</b>
Boundary Detection	Period 1	2 (0.82 – 2.52)	8 (3.96 – 12.71)
	Period 2	2 (-0.09 – 3.76)	7 (2.37 – 11.3)
Community Creation	Period 1	10 (-2.73 – 22.06)	26 (1.12 – 49.88)
	Period 2	8 (-0.64 – 16.14)	31 (6.00 – 55.00)
State-wide Comparison	Period 1	5 (2.10 – 8.57)	30 (14.64 – 46.22)
	Period 2	6 (2.76 – 9.24)	24 (9.90 – 37.10)
Ranking Analysis	Period 1	3 (-0.48 – 6.88)	9 (0.60 – 18.20)
	Period 2	3 (1.56 – 5.24)	9 (.051 – 17.50)
Municipality Analysis	Period 1	6 (0.96 – 11.04)	16 (10.67 – 20.93)
	Period 2	5 (2.48 – 5.92)	18 (4.42 – 31.18)

Figure 6-3 shows visual representations of the mean times shown previously in Table 6-1.



**Figure 6-3: Mean time per task per period for SOVAT and SPSS-GIS**

Table 6-2 also shows the mean results for time, but the results are shown by task only. Across all tasks and periods, the mean times are shorter for SOVAT than SPSS-GIS. The results

show clearly that when using SOVAT, the participants took much less time to complete the tasks than when using the combination of SPSS and GIS.

**Table 6-2: Mean time (rounded to the nearest minute) and 99% CI per task.**

<b>Task</b>	<b>SOVAT</b>	<b>SPSS-GIS</b>
Boundary Detection	2 (0.97 – 2.53)	8 (5.18 – 9.99)
Community Creation	9 (2.8 – 15.00)	29 (15.80 – 41.21)
State-wide Comparison	6 (3.92 – 7.47)	27 (18.56 – 35.91)
Ranking Analysis	3 (1.93 – 4.68)	9 ( 5.12 – 13.28)
Municipality Analysis	5 (3.08 – 7.12)	17 (11.91 – 21.69)

#### **6.4.1.2. Success Rate**

Table 6-3 shows the success rates for the study. The success rate is equal to the number of tasks answered correctly divided by the number of tasks attempted. For SOVAT, all tasks were attempted. For SPSS-GIS, some of the participants did not attempt all of the 5 tasks. Two participants attempted only 3 of the 5 tasks meaning that 4 tasks total among all of the 13 participants were not attempted. The main reason for not attempting all of the tasks was the lengthy session time for SPSS-GIS. The training session took about an hour and a half and as indicated by the previous times, some of the tasks could take a half an hour to complete. The participants were instructed that the sessions would take two and a half hours and thus some had personal appointments which prohibited them from attempting all 5 tasks.

**Table 6-3: Success rate for the tasks.**

		<b>SOVAT</b>	<b>SPSS-GIS</b>	<b>Higher Success Rate</b>
Boundary Detection	Period 1	.83	.86	SPSS-GIS
	Period 2	1.0	.50	SOVAT
Community Creation	Period 1	.67	.29	SOVAT
	Period 2	.71	.17	SOVAT
State-wide Comparison	Period 1	.83	.14	SOVAT
	Period 2	.86	.33	SOVAT
Ranking Analysis	Period 1	1.0	.83	SOVAT
	Period 2	1.0	.80	SOVAT
Municipality Analysis	Period 1	1.0	.33	SOVAT
	Period 2	.86	.40	SOVAT

The rates show that participants were more accurate using SOVAT than SPSS-GIS for all but one of the periods. The exception is the Boundary Detection task during period 1. This task was considered the easiest task by the researcher in the sense that it required the fewest steps to complete. Thus it was anticipated that this task would get the most successful results for SPSS-GIS.

Even though all of the participants had experience using SPSS, most of them commented that they “had never used SPSS for this purpose”. Hence their use of SPSS focused on statistical analysis functions, rather than the tasks required for numerical-spatial problem solving (such as data aggregation and community creation).

The community creation task and the state-wide comparison were the most difficult tasks to perform using SPSS-GIS. Many participants had difficulty creating customized communities in SPSS since they had never used SPSS for this purpose before. In addition, the participants also had difficulty computing the state-wide average amongst a very finely grained data set (with many rows of data).

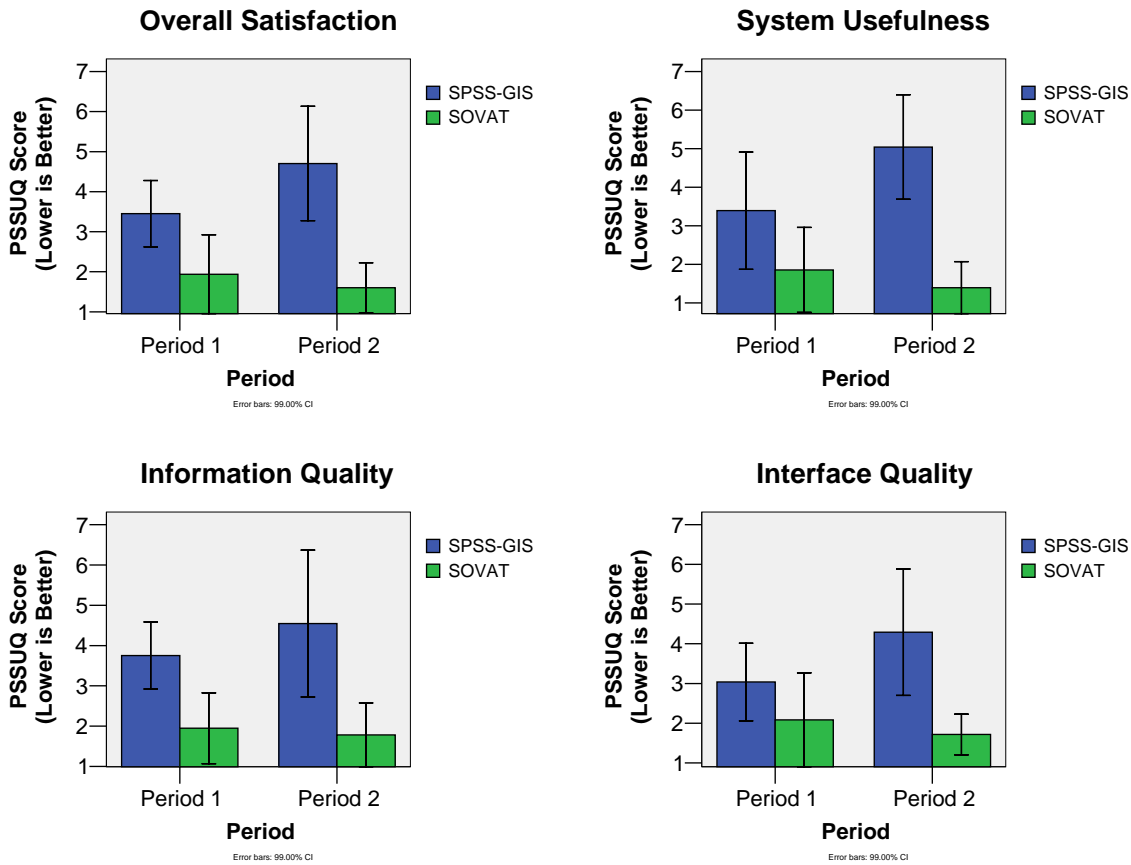
#### 6.4.2. Subjective Measurements

Table 6-4 shows the mean and 99% confidence intervals for the PSSUQ survey (with overall, as well as the satisfaction categories previously explained in chapter 4) by period. As mentioned, lower scores indicate higher levels of satisfaction. Figure 6-4 shows the same results in a bar chart format, while Table 6-5 shows the satisfaction results without breaking it down by period.

**Table 6-4: Mean Satisfaction scores and 99% CI per period.**

		<b>SOVAT</b>	<b>SPSS-GIS</b>
Overall Satisfaction	Period 1	1.94 (0.95 – 2.92)	3.45 (2.62 – 4.28)
	Period 2	1.60 (0.98 – 2.23)	4.70 (3.27 – 6.13)
System Usefulness	Period 1	1.86 (0.76 – 2.96)	3.39 (1.87 – 4.91)
	Period 2	1.40 (0.72 – 2.07)	5.04 (3.69 – 6.40)
Information Quality	Period 1	1.95 (1.07 – 2.83)	3.75 (2.92 – 4.59)
	Period 2	1.78 (0.99 – 2.58)	4.55 (2.72 – 6.37)
Interface Quality	Period 1	2.08 (0.90 – 3.27)	3.04 (2.06 – 4.02)
	Period 2	1.71 (1.20 – 2.23)	4.30 (2.70 -5.88)

Note: Lower numbers indicate higher levels of satisfaction (1= Very Satisfied, 7 = Very Dissatisfied)



**Figure 6-4: Satisfaction scores by period (A lower number is better).**

**Table 6-5: Mean Satisfaction scores and 99% CI.**

	<b>SOVAT</b>	<b>SPSS-GIS</b>
Overall Satisfaction	1.76 (1.31 – 2.20)	4.03 (3.22 – 4.84)
System Usefulness	1.61 (1.10 – 2.12)	4.16 (3.08 – 5.23)
Information Quality	1.86 (1.41 – 2.31)	4.12 (3.34 – 4.90)
Interface Quality	1.89 (1.41 – 2.36)	3.62 (2.74 – 4.49)

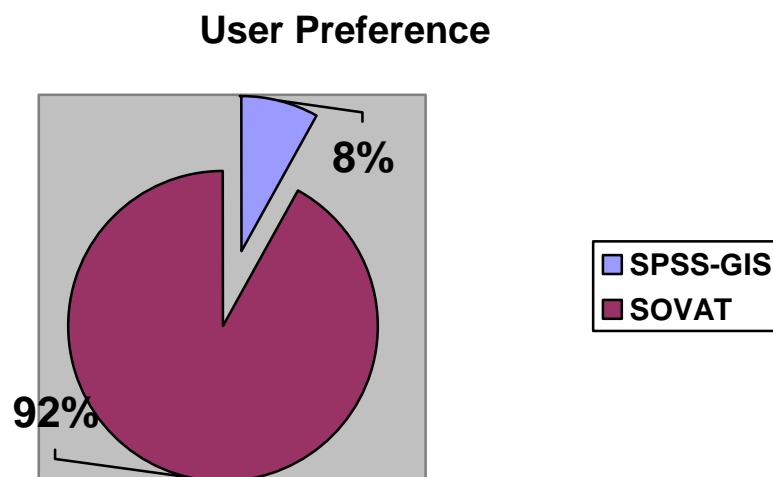
Note: Lower numbers indicate higher levels of satisfaction (1= Very Satisfied, 7 = Very Dissatisfied)

The subjective data shows that SOVAT is perceived as more satisfactory across all periods and satisfaction categories than SPSS-GIS. Within the three specific categories, system usefulness showed the greatest mean difference, while interface quality had the smallest mean difference. Since participants were asked to think of the combined use of SPSS and GIS during

completion of the questionnaire, perhaps their satisfaction with the interface of ArcView attributed to the lower score for SPSS-GIS in relation to interface quality.

#### **6.4.2.1. User Preference**

After completing the second session, before they left, a ‘mini interview’ was conducted to assess user preference. The purpose of this interview was to go beyond the numeric responses from the satisfaction questionnaire and obtain more qualitative feedback regarding their attitudes towards both systems. The first question that was posed to every individual was “What system did you like better and why”? The responses are shown in the pie chart (Figure 6-5). In total, 12 of the 13 participants (92%) preferred SOVAT, while 1 of the 13 participants (8%) preferred the combination of SPSS and GIS. A chi-square test shows statistical significance between user preferences of the two different systems ( $p = .002$ ).



**Figure 6-5: User preference between SOVAT and SPSS-GIS**

The individual responses from the post-study interview were then categorized into groups. A participant could have more than one response if they commented on more than one aspect of

the system. The counts for these groups and some example responses are shown in Table 6-6. For example one participant could mention that SOVAT “took less steps”. This response was grouped into the *Easier To Use* Category. If they then also indicated that the “SOVAT interface looked better”, then this response was grouped into the *Interface* category.

**Table 6-6: Positive responses in relation to SOVAT during the post-study interview.**

<b>Reason</b>	<b>Number of Participants who gave this Reason</b>	<b>Participant Comments in Relation to this Reason</b>
Easier to Use	12	<ul style="list-style-type: none"> <li>• “Streamlined for this purpose” [numerical-spatial problem solving]</li> <li>• “Took less steps” [to solve tasks]</li> <li>• “Easier to go back” [if a mistake is made]</li> <li>• “Very straightforward”</li> <li>• “Not as complicated”</li> </ul>
Interface	6	<ul style="list-style-type: none"> <li>• Everything was together [numerical and spatial information]</li> <li>• “1 program vs. 2”</li> <li>• “Loved the interface; the layout; organized nicely; visually appealing”</li> <li>• “Layout was well designed”</li> <li>• “SOVAT interface looked better”</li> </ul>
Information Access	4	<ul style="list-style-type: none"> <li>• “Gave you the answer quickly”</li> <li>• “Easy to get information”</li> <li>• “Easier to find information”</li> <li>• “Finding data was easier”</li> </ul>
Specific Features	6	<ul style="list-style-type: none"> <li>• “Liked Search Boxes”</li> <li>• “Drill-out and community creation [spatial modeling]”</li> <li>• “Easy to create communities”</li> <li>• Drill-out helped for boundary detection”</li> <li>• “Rates already provided”</li> </ul>

As can be seen, the majority of the positive responses towards SOVAT were in relation to its ease of use. For numerical-spatial problem solving, many of the participants felt that it was “streamlined for this purpose”. This notion of ease of use is most related to “System Usefulness”

within the PSSUQ questionnaire. Thus the particularly large mean difference for System Usefulness (Table 6-5) between SOVAT and SPSS-GIS is supported through these responses. The participants also like the layout and design of the SOVAT interface better than SPSS-GIS. The most popular response in relation to this theme was that they like “1 program vs. 2”. Hence, having to go back and forth between numerical and spatial data displays was not as popular and the combined numerical-spatial interface of SOVAT. Finally, the special features that distinguish SOVAT from other systems were also popular. For example, it was described how one of the most difficult tasks to perform using SPSS-GIS was the community creation task. One of the popular features in SOVAT was the facility of creating customized communities. Another feature that was popular was the drill-out function, which participants indicated “helped for boundary detection”. As mentioned in section 3.1, drill-out is specific only to SOVAT and is not available in other OLAP-GIS systems.

Table 6-7 shows the negative responses towards SOVAT. As described earlier in section 6.4.2, the interface category within the satisfaction questionnaire showed the tightest gap between mean scores of SOVAT and SPSS-GIS. It was believed that the interface of ArcGIS was thought fondly of and contributed to the better scores for SPSS-GIS for this category. There were also some features about the SOVAT interface that received some negative feedback such as: “the bar chart was always changing color”. This was discovered during the usability study but unfortunately can not be addressed based on the nature of the presentation layer components used within the system.

**Table 6-7: Negative responses in relation to SOVAT during the post-study interview.**

<b>Reason</b>	<b>Number of Participants who gave this Reason</b>	<b>Participant Comments in Relation to this Reason</b>
Interface	4	<ul style="list-style-type: none"><li>• “Default setting. Allegheny was always darker”</li><li>• “The bar chart was always changing color”</li><li>• “No option to ‘sort’ bars in bar chart”</li><li>• “Map was not easy to navigate at Municipality level”.</li></ul>
Information Access	1	<ul style="list-style-type: none"><li>• “Not as comprehensive as SPSS.”</li><li>• “Difficult to find information on screen”</li></ul>

### **6.4.3. Mixed Model Analysis**

#### **6.4.3.1. Time**

The mixed model analysis was used for obtaining inferential results for time and user satisfaction. The mixed model extends on the general linear model (GLM) to allow for fixed (treatment, period, group) and random effects (subjects) [102]. Both fixed and random variables are present in crossover designs like the one outlined in this chapter and thus it was decided to use this model for inferential purposes. Table 6-8 shows the p-values for the three different effects in the study: *group* or sequence (SOVAT → SPSS-GIS, or SPSS-GIS → SOVAT), *period* (period 1, period 2), and *intervention* (SOVAT, SPSS-GIS).

**Table 6-8: Mixed model analysis of Time variable. Shown are p-values per effect per task.**

	Boundary Detection	Community Creation	State-wide Comparison	Ranking Analysis	Municipality Analysis
<b>Group</b>	.338	.465	.250	.833	.269
<b>Period</b>	.429	.742	.244	.944	.953
<b>Intervention</b>	.000	.001	.000	.001	.000

At the .01 level, there is no group effect for any of the tasks. This indicates that the participants were sufficiently randomized to each group in relation to the variable *time*. The p-values for period indicate that there is no period effect for time. This indicates that the period (1 or 2) does not effect the time to complete the tasks. The intervention effect is significant at the .01 level. This indicates that the type of system used (SOVAT or SPSS-GIS) impacted the time to complete the tasks. As can be seen from the descriptive results, the participants completed the tasks in much shorter time than when they used a combination of SPSS and GIS.

#### 6.4.3.2. User Satisfaction

The mixed model results for user satisfaction are shown in Table 6-9.

**Table 6-9: Mixed model analysis of User Satisfaction.**

	Overall Satisfaction	System Usefulness	Information Quality	Interface Quality
<b>Group</b>	.004	.003	.108	.008
<b>Period</b>	.080	.072	.283	.125
<b>Intervention</b>	.000	.000	.000	.000

The group effect corresponds to the treatment\*period interaction which is an alias for a carryover effect [103]. As can be seen, the group effect is significant at the .01 level for overall satisfaction, system usefulness, and interface quality. It is not significant for information quality ( $p = .108$ ). This indicates that a carryover effect was present in relation to participant responses on the satisfaction questionnaire. There are many possible reasons for this. One may be the similarity of the tasks (they are similar, but not the same). That is, the participant uses SOVAT in period 1 and sees similar tasks for period 2. The participant believes that they can easily complete these types of tasks but then has difficulty during the session using SPSS-GIS. Another reason is that SOVAT is so unique, and thus the participants have not used anything like

it before for numerical-spatial problem solving. Participants who use SPSS-GIS first give 'ok' or luke-warm responses towards the system; whereas a participant who used SOVAT first and SPSS-GIS second now realizes that there is a system that is so much better for these types of problem solving tasks. The natural response is to have negative sentiment towards the inferior system. As the charts in Figure 6-4 indicate, SOVAT is always better in period 2, while SPSS-GIS is always worse in period 2. This is consistent with the belief that SOVAT is perceived as better than SPSS-GIS. The introduction of the novel system (SOVAT) for difficult numerical-spatial tasks will impact the satisfaction level of the old system (SPSS-GIS) and vice-versa. A similar comparison can be drawn between UNIX and Windows operating systems. A long-time UNIX user will likely give high satisfaction scores to Windows in regards to difficult tasks that they used to do in UNIX. If they are then asked to use UNIX to solve these similar difficult tasks, then their satisfaction towards UNIX will decrease based on their recent experience with Windows. Conversely, if they use Windows after using UNIX, then they are likely to recall their recent difficulties using UNIX for similar tasks and thus give higher satisfaction scores to Windows.

In relation to period effect, there is no significant effect at the .01 level, suggesting that the period does not influence the satisfaction level of the participant. The treatment effect is significant at the .01 level for all satisfaction categories. This supports the mean results from the satisfaction questionnaire that suggest that the participants are more satisfied with SOVAT than with SPSS-GIS.

## 6.5. Discussion

Both the descriptive results and the inferential results favored SOVAT over the combination of SPSS and GIS for numerical-spatial tasks. The difference in time was substantial for all of the 5 tasks. This was anticipated before the study began. As explained in chapter 2, the characteristics of OLAP allow for very fast retrieval of data within the multidimensional cube. Thus, any retrieval with SOVAT would take seconds. In addition, the enhancement of the interface as a result of the rigorous usability study (chapter 4) tailored the design to meet the needs of numerical-spatial problem solving. On the other side, SPSS-GIS requires many steps in order to complete a typical numerical-spatial task. A typical scenario might involve the participant needing to:

1. Open ArcGIS and manually identify bordering areas on a map.
2. Open SPSS and attempt to find a diagnosis among thousands of rows (or cases) of data
3. Type a complex “Select Cases” command that requires a statement to be syntactically accurate.
4. Aggregate the selected cases by choosing appropriate break (or grouping) variables as well as the numerical measure on which to sum.
5. Calculate the numerical rates.
6. Return to the large SPSS file and specify a subset of the original selected cases
7. Aggregate the smaller subset of cases by selecting a different break variable and the numerical measures to sum.
8. Calculate a new rate based on this latest aggregation.

Errors can be easily made within any of these problem solving steps. The most critical errors during the study were ones in which the participant needed to retreat back to a previous

step. For example, one error that occurred was when a participant in performing step 3, typed in a long “selected cases” statement incorrectly. Syntactically, the statement was correct. However, the participant did not realize until step 5 that they forgot to include a county name in the statement. They thus needed to go back to the original file and retype the select cases statement. Many other examples like this occurred in the study and contributed to the long task completion times for SPSS-GIS. These scenarios also added to user frustration and contributed to poorer satisfaction scores for SPSS-GIS.

The same scenario in SOVAT might require a few simple mouse clicks while providing the participant the ability to add elements (such as a forgotten county) at any time without having to retrace steps. This notion of “error forgiveness” became a big theme of the evaluation study and was one of the main ingredients in widening the gap between the two systems in relation to the objective and subjective results.

## **6.6. Limitations**

The omission of intended users from this evaluation was a limitation to the study. It would have been preferred to use actual professionals in academic settings, health departments, or government agencies who use a combination of SPSS and GIS for numerical-spatial problem solving tasks. This would have strengthened the generalizability of the results to include working community health professionals. Unfortunately, as was the case in the usability study, time away from work, and difficulty in identifying local professionals who meet the specific participation requirement, did not allow for this population to be sampled. It was believed that obtaining potential future users of the system, such as graduate students, researchers, and some faculty within the health sciences programs at the University of Pittsburgh was the best

alternative. It is believed that this population provided excellent results on which to compare the two independent variables.

## **6.7. Conclusion**

This within-subjects crossover summative evaluation compared SOVAT to SPSS-GIS for typical numerical-spatial problem solving in community health assessment analysis. Thirteen graduate students, faculty, or researchers in the health sciences with prior SPSS experience were recruited for the study. Half were randomly assigned to use SPSS-GIS first, while the other half were randomly assigned to use SOVAT first. Regardless of the system, the participants received a training video and completed two practice tasks using the system. They then completed 5 real tasks that were similar in nature to the practice tasks. Time and answer were analyzed using screen capture software. A post study satisfaction questionnaire was completed after each session. A 2 – 3 week interlude period existed between the two sessions. For the second session, the participants were asked to use the system that they did not use during period 1. As in the first session, they received a training video and two practice tasks before attempting the 5 actual study tasks. Time and answer were measured using screen capture software, while user satisfaction was recorded using a valid and reliable questionnaire. Before leaving the second session, a “mini interview” session took place in which the participants were asked which system they preferred. This allowed for more open-ended responses in relation to satisfaction that were not provided on the numerical-based questionnaire.

Time, answer, and satisfaction were analyzed using descriptive statistics (mean and CI) while time and satisfaction were also analyzed using inferential statistics (linear mixed model analysis). The results for time all favored SOVAT to the combination of SPSS and GIS across

all periods and tasks. This was shown at the 99% confidence interval. In addition SOVAT was superior to SPSS-GIS for answer at the 99% confidence interval for all tasks and periods, indicating that using SOVAT for these tasks led to more accurate responses. Finally SOVAT was superior to SPSS-GIS for all tasks in relation to all satisfaction categories (overall, system usefulness, information quality, and interface quality). This was also at the 99% CI level.

The mixed model analysis provided p-values for the different effects of the crossover design: group (or sequence), period, and treatment. For time, across the 5 tasks, group ( $p = .250 - .833$ ) and period ( $p = .244 - .944$ ) were not significant. This indicates that the ordering of exposure to the independent variables as well as the progress from period 1 to period 2 did not influence their time in completing the tasks. The treatment effect comparing SOVAT to SPSS-GIS was significant for time at the .01 level ( $p \sim .000$  for all 5 tasks). This suggests that SOVAT was more efficient to use in solving the tasks.

Satisfaction was measured for overall satisfaction (all 19 questions) and for the specific satisfaction categories on the questionnaire (system usefulness, information quality, interface quality). Group was shown to be significant at the .01 level for overall satisfaction ( $p = .004$ ), system usefulness ( $p = .003$ ), and interface quality ( $p = .008$ ). This suggests the presence of a carryover effect between the groups. Reasons for this attributed to the similarity between the tasks of the two sessions and to the fact that SOVAT was so unique and novel for these types of problem solving situations. Period ( $p = .072 - .283$ ) was not significant for any of the satisfaction categories at the .01 level indicating that the progress from period 1 to period 2 did not influence their satisfaction towards either system. The treatment effect was significant ( $p = .000$ ) indicating that SOVAT was perceived as more satisfactory across all categories when compared to SPSS-GIS.

The responses from the mini post-study interview enforce the notion that SOVAT was perceived as more satisfactory than SPSS-GIS for these tasks. The question “Which system did you like better and why?” was posed to all participants. Chi-square analysis shows that there is a statistically significant difference between the number of people who preferred SOVAT (92%) to SPSS-GIS (8%) ( $p = .002$ ). Responses from the interview were categorized into groups. The most popular reason of why people preferred SOVAT was in relation to ease of use. Example responses included: “Everything was together [numerical and spatial information]” and “Very straightforward”. Additional categories included interface reasons, and the special features that make SOVAT unique (drill-out for boundary detection, community creation, etc.). The most telling response that SOVAT was considered better than SPSS-GIS was from one participant who summed it all up by saying, “This was fabulous”.

## **7. SUMMARY AND FUTURE DIRECTIONS**

### **7.1. Summary**

The purpose of this research was to examine the potential for an OLAP-GIS system in relation to numerical-spatial problem solving. The Spatial OLAP Visualization and Analysis Tool (SOVAT) was developed for this purpose. SOVAT is unique in that it combines On-Line Analytical Processing (OLAP) with Geospatial Information System (GIS) capabilities. A system like this has never been adequately evaluated for numerical-spatial problem solving in community health assessment analysis.

The major questions introduced by this study are:

1. What are the usability issues associated with a novel OLAP-GIS interface that combines numerical and spatial information?
2. What are the current types of information technology used by professionals for numerical-spatial problem solving in community health?
3. How does an OLAP-GIS system compare to existing information technology used by today's professionals for numerical-spatial problem solving?

I have addressed these questions by conducting the following studies:

1. Usability Study - This identified several issues in relation to human-computer interaction during numerical-spatial problem solving. These issues allowed for a better and more usable OLAP-GIS system to be developed.
2. Community Health Assessment Survey – This identified the current information technology used by community health assessment professionals for numerical-spatial problem solving. The online survey broke numerical-spatial tasks into steps which

provided specific detail of how the individual software applications are used together for solving these types of tasks.

3. Summative Comparative Evaluation - This enabled for statistical hypothesis to be supported through inferential and descriptive analysis. The crossover design allowed for participants to provide objective and subjective feedback for both systems (SOVAT, and the current technology, SPSS-GIS).

These studies provide a starting point for the analysis of OLAP-GIS systems for numerical-spatial problem solving. Many questions remain unanswered and need to be explored further.

## **7.2. Future Directions**

### **7.2.1. OLAP-GIS for other domains**

It is believed that SOVAT can be applied to many different domains outside of community health for the purposes of numerical-spatial problem solving. I would thus like to explore the impact of an OLAP-GIS system in other public health-related fields such as environmental health and cancer.

Environmental health professionals have begun to use GIS for their analysis of environmental health factors (for some examples, see [104-110]). Geospatial technology lends itself to this type of analysis. I am interested in seeing how SOVAT can be applied to environmental health problem solving. Most likely a rigorous usability study would need to be performed in order to identify how environmental health professionals perceive an OLAP-GIS interface. The visual interface expectations for an environmental health professional are seemingly much different than a community health professional's expectations. Different geospatial layers are required as well as different interface functions. It is likely that an

individual examining environmental health data would need a more detailed map that contains specific landmarks, waterways, and buildings. This might not be the case with a community health professional where the spatial needs stop at the municipality level without the emphasis on specific city markers. In addition, it is probable that an environmental health expert would need specific spatial functions such as buffering and shortest path analysis. How these features and functions are perceived in relation to an OLAP-GIS interface would need to be analyzed through rigorous usability analysis. A similar study progression as this dissertation could be conducted that then uses a survey to identify the current information technology for environmental health research and then compares these current applications to an OLAP-GIS system during a summative comparative evaluation.

#### **7.2.2. Cognitive Issues in relation to an OLAP-GIS system**

OLAP and GIS are routinely discussed and analyzed in a very complex quantitative manner. It is important to bring these technologies outside this complex area and analyzed for a more qualitative and cognitive manner. As mentioned in chapter 6, Hart [99] has been one of the few individuals who has studied cognitive variables in relation to OLAP. These issues go beyond Davis's Technology Acceptance Model (TAM) [100] and include additional variables of interest. The novelty of an OLAP-GIS system makes this type of exploration very desirable. It would also be interesting to explore the differences between OLAP-GIS and OLAP, or OLAP-GIS and GIS, across different cognitive variables. This could be done by allowing participants to use SOVAT for a few weeks and then giving them a survey that allows for the recording of different cognitive variables. The same could be done for GIS and for OLAP separately enabling for the results to be compared.

# **GSPH Grad Students Needed**

## **Usability of a Decision Support System**

You will be asked to use a decision support system to complete assigned tasks, and to fill out a background survey and two questionnaires. All will be finished in an hour and thirty minutes and you will receive **\$30** as compensation.

The whole experiment will be conducted on campus at 8084 Forbes Tower (Center for Biomedical Informatics Conference Room).

Contact Matthew Scotch at (412) 647-7306 (email: [scotch@cbmi.pitt.edu](mailto:scotch@cbmi.pitt.edu)) for more information or for participation. Please leave your name and a phone number at which you can be contacted for an appointment.

## APPENDIX B: Recruitment Letter for Usability Study

Dear X,

You have been identified an active and knowledgeable individual in the field of community/public health assessment research. We have recently developed a system, Spatial OLAP Visualization and Analysis Tool or SOVAT, which is intended to be used as a decision support system for this type of research.

The purpose of this research is to identify human-computer usability issues associated with our SOVAT system user interface and thus we are evaluating the system and not you. This study will last approximately 1 hour and a half and will take place in the conference room in the Center for Biomedical Informatics on the campus of the University of Pittsburgh. There will only be one researcher present in the room. The conference room door will be closed for privacy. The evaluation process will be audio recorded. We will also use screen-capture software to record you interaction with the computer.

You will be given tasks and asked to complete them using our system (our system will be demonstrated to you before the evaluation begins). During this process you will be requested to ‘Think-Aloud’ or verbalize your thoughts (information and instruction of this process will be provided during the study). There will be 5 individual tasks. After completion of all the tasks, you will be asked to complete 2 short questionnaires aimed at identifying the usability of our system.

There are no foreseeable risks associated with this project, nor are there any direct benefits to you. Each participant will receive \$30 as a token of our appreciation.

This is an entirely anonymous study, and so your audio feedback during human-computer interaction, as well as your written responses during the after study questionnaire, will not be identifiable in any way. Responses are confidential and results will be kept under lock and key. Your participation is voluntary, and you may withdraw from the project at any time.

Please feel free to contact me at (412) 647-7306 or [scotch@cbmi.pitt.edu](mailto:scotch@cbmi.pitt.edu), if you have any questions or would like to enroll in the study. If you choose to enroll, you will be contacted via telephone to establish an appropriate study time.

Sincerely,

Matthew Scotch, MA  
Doctoral Candidate

## **APPENDIX C: Informed Consent Form for Usability Study**

Current Approval Date: February 1, 2005  
Modification Approval Date: May 24, 2005  
Renewal Date: January 31, 2006  
University of Pittsburgh  
Institutional Review Board  
IRB# 0501060

### **SPATIAL OLAP VISUALIZATION AND ANALYSIS Tool (SOVAT)** **USABILITY STUDY**

#### **CONSENT TO ACT AS A PARTICIPANT IN A USABILITY STUDY**

**TITLE:** SPATIAL OLAP VISUALIZATION AND ANALYSIS TOOL (SOVAT) Usability Study

**PRINCIPAL INVESTIGATOR:**

Matthew Scotch, M.A.  
Ph.D. Candidate  
Center for Biomedical Informatics  
University of Pittsburgh  
8084 Forbes Tower  
Telephone: 412-647-7306

**CO-INVESTIGATORS:**

Bambang Parmanto, Ph.D.  
Assistant Professor  
Center for Biomedical Informatics  
University of Pittsburgh  
8084 Forbes Tower  
Telephone: 412-383-6649

Valerie Monaco, Ph.D.  
Assistant Professor  
Center for Biomedical Informatics  
University of Pittsburgh  
8084 Forbes Tower  
Telephone: 412-647-3064

Ravi K. Sharma, Ph.D.  
Assistant Professor  
Department of Behavioral and Community Health Sciences  
Graduate School of Public Health  
University of Pittsburgh  
PUBHL 228  
Telephone: 412-624-3615

**SOURCE OF SUPPORT:** None

***What is the purpose of this usability study?***

Current decision support systems are perceived as inadequate for spatial and numerical problem solving. We have developed a decision support system that enhances the development of spatial and numerical scenarios (queries) for solving these types of problems.

You are being asked to participate in a research study in which we will test the usability of our SOVAT system during human-computer interaction.

It is anticipated that SOVAT will enhance spatial and numerical problem solving.

In this research study, we will examine the usability issues associated with our system during Human-Computer Interaction (HCI) using the field of community health assessment as our case study. In addition, you will also fill out 2 questionnaires about your satisfaction with the system.

***Who is being asked to take part in this research study?***

You are being asked to take part in this research study because you are a potential or active professional researcher in the field of Community Health Assessment. The study is being performed on a total of 6-12 individuals.

Males and females age 18 years or older are being recruited for this study.

***What will my participation in this research study involve?***

If you agree to participate in the usability study of our SOVAT system, you will be asked to perform the following procedures:

**Experimental Procedures**

1. Complete a short computer background questionnaire
2. You will be given instructions and practice in “thinking-aloud”. You will be asked to think aloud while you use our system.
3. You will be given a series of 5 tasks to complete using our SOVAT system via an external mouse. You will be asked to think aloud while you try to complete the task. The actions you take with the mouse will be recorded via screen capture software, and the vocal responses you elicit will be captured via an audio recorder.
4. Following the completion of all the tasks, you will be given 2 surveys about your overall experience using the system.
5. The entire session should take between 60 and 90 minutes to complete

***What are the possible risks, side effects, and discomforts of this research study?***

There is no risk of physical injury associated with your participation in the usability test. Since we don't collect your personal identifiable information, participation in the study is not expected to involve the possible risk that your information is known to other individuals, although there always exists a risk for breach of confidentiality. The risk is minimized by keeping your research information confidential.

***Will I be paid if I take part in this research study?***

You will be given \$30. In addition, any parking fees related to your participation in this study will be paid for by the study.

***Who will know about my participation in this research study?***

All records related to your involvement in this research study will be stored via a privately password-protected file on one of the researcher's computers. Your identity on these records will be indicated by a unique identification number rather than by your name, and the information linking these numbers with your identity will be kept separate from the research records (also password protected). Only the researchers listed on the first page of this page of this form and their staff will have access to your research records. Your research records will be destroyed when such is approved by the sponsor of this study or, as per University policy, at 5 years following study completion, whichever should occur first.

Any information about you obtained from this research will be kept as confidential (private) as possible. You will not be identified by name in any publication of research results unless you sign a separate form giving your permission (release). In unusual cases, your research records may be released in response to an order from a court of law.

***Who will have access to identifiable information related to my participation in this research study?***

In addition to the investigators listed on the first page of this consent form and their research staff, the following individuals will or may have access to identifiable information related to your participation in this research study:

Authorized representatives of the University of Pittsburgh Research Conduct and Compliance Office may review your identifiable research information for the purpose of monitoring the appropriate conduct of this research study. In unusual cases, the investigators may be required to release identifiable information related to your participation in this research study in response to an order from a court of law. If the investigators learn that you or someone with whom you are involved is in serious danger or potential harm, they will need to inform, as required by Pennsylvania law, the appropriate agencies.

***Is my participation in this research study voluntary?***

Your participation in this research study is completed voluntary. You do not have to take part in this research study and, should you change your mind, you can withdraw from the study at any



Current Approval Date: February 1, 2005  
Modification Approval Date: May 24, 2005  
Renewal Date: January 31, 2006  
University of Pittsburgh  
Institutional Review Board  
IRB# 0501060

**CERTIFICATION of INFORMED CONSENT**

I certify that I have explained the nature and purpose of this research study to the above-named individual(s), and I have discussed the potential benefits and possible risks of study participation. Any questions of the individual(s) have about this study have been answered, and we will always be available to address future questions as they arise.

---

Printed Name of Person Obtaining Consent Role in Research Study

---

Signature of Person Obtaining Consent

---

Date

## APPENDIX D: Computer Background Questionnaire

### I. Demographics

- a. Your age: \_\_\_\_\_
- b. Your gender: ☐ Female ☐ Male

### II. Computer Experience

- a. In a typical week, how many hours do you personally use a computer hands-on?
- \_\_\_\_\_

**If you answered zero, go to question e.**

- b. What kind(s) of computer(s) do you use? (Check all that apply)
- ☐ Macintosh
- ☐ IBM PC or compatible
- ☐ Terminal connected to a remote mainframe computer (e.g. hospital information system)
- ☐ High-performance scientific workstation
- ☐ Other (explain) \_\_\_\_\_
- c. To what extent do you personally use a computer for each of the following professional tasks? Please circle your answer.

1. **Never** perform this task.

2. **Perform** this task but **never** use a computer.

3. **Sometimes** use a computer

4. **Often** use a computer

5. **Always** use a computer

	1	2	3	4	5
Documenting information					
Accessing information					
Communicating with colleagues					
Writing (reports, research papers, teaching materials)					
Preparing presentation slides or overheads					
Performing statistical analysis on health data					
Online Search Engines (Google, Yahoo!)					

- d. What kind(s) of computer(s) do you routinely use? (Check all that apply)
- ☐ Desktop computer at your office
- ☐ Desktop computer at home

- ☐ Portable or notebook computer
  - ☐ Other (please specify: \_\_\_\_\_)
- e. What training or experience with computers have you had? (check all that apply)
- ☐ Formal undergraduate course(s) in computer science or related field
  - ☐ Formal graduate course(s) in computer science or related field
  - ☐ Workshops on conference on computers
  - ☐ Self-guided learning on computers
  - ☐ None
- f. On the whole, how sophisticated a computer user do you consider yourself?
- ☐ Very sophisticated
  - ☐ Sophisticated
  - ☐ Neither sophisticated nor unsophisticated
  - ☐ Unsophisticated
  - ☐ Very unsophisticated

*Note: This questionnaire was adapted from the survey "Computers in Medical Care" developed by Charles Friedman, PhD and William Detmer, MD, MSc in 1993 and published in: Cork RD, Detmer WM, Friedman CP. "Development and initial validation of an instrument to measure physicians' use of, knowledge about, and attitudes toward computers". JAMIA 1998. 5(2): 164.*

## **APPENDIX E: Tasks for Usability Study**

### **Usability Tasks**

#### **Task 1**

*How does the death rate per 100,000 of Allegheny County in 1996 compare to the death rate of the counties that border it?*

## Task 2

*For this task Northwestern PA is defined by the following counties: Crawford, Erie, Forest, Mercer, Venango, Warren. Northeastern PA is defined by: Monroe, Pike, Susquehanna, Wayne, and Wyoming.*

*What is the difference in cancer incidence rate per 100,000 of female “Malignant Neoplasm of Pancreas” in 1998 between northwestern and northeastern PA?*

## Task 3

*How does the inpatient rate per 1,000 in 2000 of “complications occurring mainly in the course of labor and delivery” in Forest County compare to the counties that border it? For the county with the highest rate, what are the top 5 municipalities? Do all these municipalities border each other?*

## Task 4

*What are the top 5 “circulatory system” diagnoses of Inpatient rate per 1,000 for Males Aged 65-74 in McKean County in 1999? For the diagnosis with the highest rate, how does this rate compare to the state?*

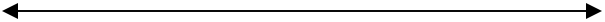
## Task 5

*What are the top 5 municipalities in Bucks County that have the highest death rate per 100,000 from “Asthma” in 2000? How do these municipalities compare to the state?*

## APPENDIX F: PSSUQ Questionnaire


### IBM Post-Study System Usability Questionnaire (PSSUQ)

1. Overall, I am satisfied with how easy it is to use this system.

strongly agree								strongly disagree	
	1	2	3	4	5	6	7		N/A


Comments:

2. It was simple to use this system.

strongly agree								strongly disagree	
	1	2	3	4	5	6	7		N/A

Comments:

3. I could effectively complete the tasks and scenarios using this system.

strongly agree								strongly disagree	
	1	2	3	4	5	6	7		N/A


Comments:

4. I was able to complete the tasks and scenarios quickly using this system.

strongly agree								strongly disagree	
	1	2	3	4	5	6	7		N/A

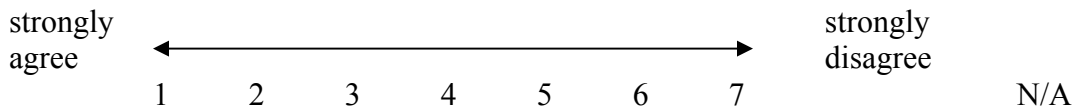
Comments:

5. I was able to efficiently complete the tasks and scenarios using the system.

strongly agree								strongly disagree	
	1	2	3	4	5	6	7		N/A

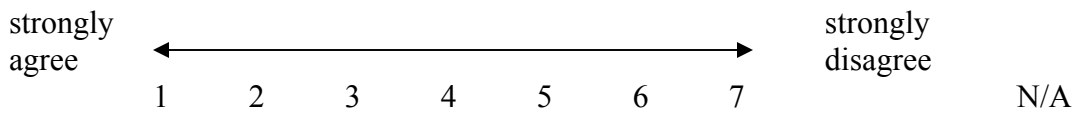
Comments:

6. I felt comfortable using this system.



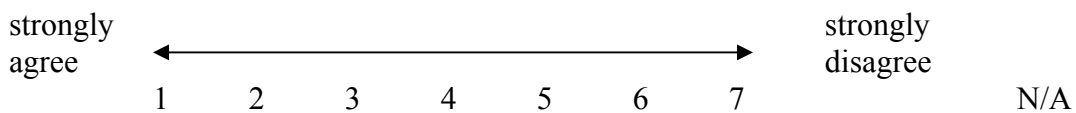
Comments:

7. It was easy to learn to use this system.



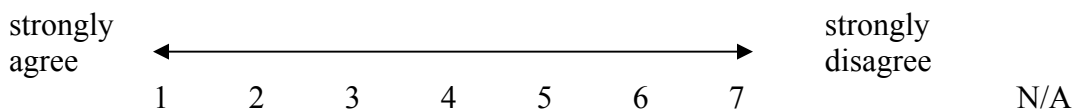
Comments:

8. I believe I could become productive quickly using this system.



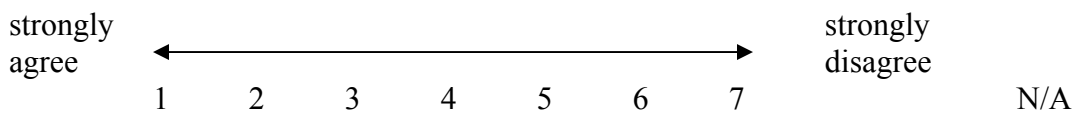
Comments:

9. The system gave error messages that clearly told me how to fix problems.



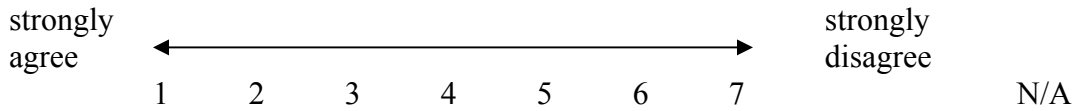
Comments:

10. Whenever I made a mistake using the system, I could recover easily and quickly.



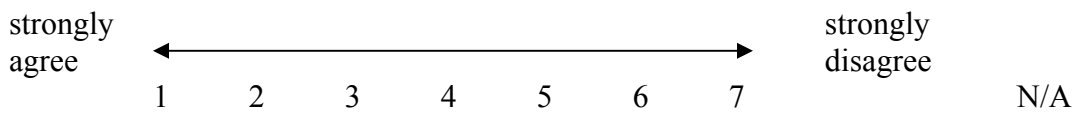
Comments:

11. The information (such as on-line help, on-screen messages, and other documentation) provided with this system was clear.



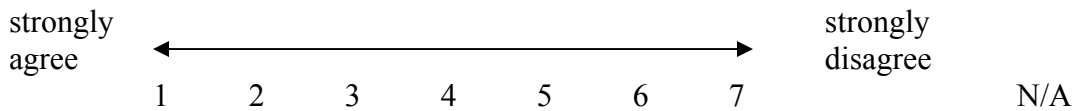
Comments:

12. It was easy to find the information I needed.



Comments:

13. The information provided for the system was easy to understand.



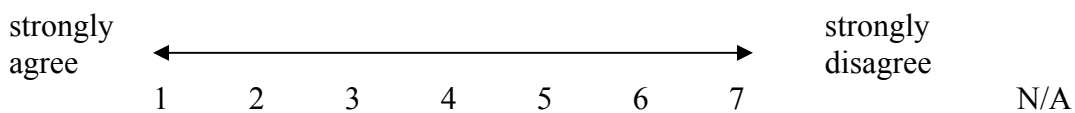
Comments:

14. The information was effective in helping me complete the tasks and scenarios.




Comments:

15. The organization of information on the system screens was clear.



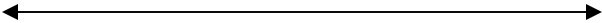
Comments:

16. The interface of the system was pleasant.

strongly agree								strongly disagree	
	1	2	3	4	5	6	7		N/A


Comments:

17. I liked using the interface of the system.

strongly agree								strongly disagree	
	1	2	3	4	5	6	7		N/A

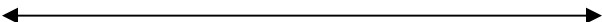
Comments:

18. This system has all the functions and capabilities I expect it to have.

strongly agree								strongly disagree	
	1	2	3	4	5	6	7		N/A

Comments:

19. Overall, I am satisfied with this system.

strongly agree								strongly disagree	
	1	2	3	4	5	6	7		N/A

Comments:

## Spatial OLAP Visualization and Analysis Tool (SOVAT) Post Session Questionnaire

What were the three worst things about this system?

Do you have any additional questions you'd like to ask about the system?

How frustrated are you feeling at this moment?

*Thank you again for your help. I'll need you to fill out this receipt for me in order to demonstrate that I paid you. Here is your payment. Over the next several months, we will attempt to analyze your session and other sessions to make improvements to the system. Thanks for helping us make it a better system*

## APPENDIX H: Best and Worst Results from Usability Study

Round 2				
<b>Worst Things</b>	1.	Did not know how to construct groups of counties without having to select them from the map	4.	Community Comparisons
	2.	Understanding Drill-Up and Drill-down	5.	Drill-down to municipalities + zooming in on a subsequent county
	3.	Default selection of all the PA counties under “Geography”	6.	Error message for community
<b>Best Things.</b>	7.	The response time is slow	8.	Mouse function is not clear
	9.	Limited screen size makes results difficult to read.		
Round 3				
<b>Worst Things</b>	1.	No ‘Help’ icon	4.	Need a ‘Help’ button
	2.	Not clear as to the directions on how to use the system	5.	It was difficult to understand it from the explanation given
	3.	The video was not helpful in providing information	6.	The messages on the screen are not clear and helpful
<b>Best Things</b>	7.	Parts of the screen are difficult to understand (ex. lower left corner)		
	1.	The ability to drag and drop	4.	It easily allowed you to manipulate the data
	2.	Speed with which data are displayed	5.	Gave you a visual of the data
			6.	It contained many attributes for your use
Round 4				
<b>Worst Things</b>	1.	When something went wrong, it didn’t offer suggestions of possible alternative solutions	4.	Got confused in then end with what was meant by ‘All Municipalities’
	2.	Sometimes the map was either too small or had insufficient resolution	5.	Labeling on the map
	3.	The easy number in the bar chart was not always easy to read (no gridlines)	6.	Changing color on the bar charts
<b>Best Things</b>	7.	Knowing how to get out of somewhere I didn’t want to be (drill-down on map)		
	1.	Overall easy to use	4.	Filters were quite easy to use
	2.	Intuitive	5.	Actual bar charts were easy

## APPENDIX I: Objective and Subjective Results from Usability Study

### Usability Study - PSSUQ Results by Round

Category	Round 1	Round 2	Round 3	Round 4	Round 5
System Usefulness	4.38	2.58	3.75	3.21	2.42
Information Quality	5.0	3.33	3.71	3.76	2.48
Interface Quality	4.58	2.75	3.0	2.58	2.17
<b>Overall</b>	<b>4.65</b>	<b>2.89</b>	<b>3.58</b>	<b>3.28</b>	<b>2.39</b>

### Usability Study - Objective Results by Round

	Usability Criteria	Goal	Round 1	Round 2	Round 3	Round 4	Round 5	Win
<b>Task 1</b>	Time (min)	<5	TNC*	9	17	6	4	Round 5
	Erroneous Action	<5	8	8	18	3	1	Round 5
	Problem Space	<3	3	2	3	0	0	Rounds 4, 5
	Answer	Correct	0 Correct 3 Wrong	3 Correct 0 Wrong	3 Correct 0 Wrong	3 Correct 0 Wrong	3 Correct 0 Wrong	Rounds 2,3,4,5
<b>Task 2</b>	Time (min)	< 5	TNC	14	16	9	5	Round 5
	Erroneous Action	<5	12	14	12	3	1	Round 5
	Problem Space	<3	7	6	4	0	0	Rounds 4,5
	Answer	Correct	1 Correct 2 Wrong	2 Correct 1 Wrong	3 Correct 0 Wrong	2 Correct 1 Wrong	3 Correct 0 Wrong	Rounds 3,5
<b>Task 3</b>	Time (min)	<5	TNC	12	TNC	14	7	Round 5
	Erroneous Action	<5	14	10	14	10	2	Round 5
	Problem Space	<3	6	3	3	1	1	Rounds 4,5
	Answer	Correct	0 Correct 3 Wrong	0 Correct 3 Wrong	1 Correct 2 Wrong	2 Correct 1 Wrong	2 Correct 1 Wrong	Rounds 4,5
<b>Task 4</b>	Time (min)	<5	TNC	8	TNC	12	12	Round 2
	Erroneous Action	<5	9	1	4	2	4	Round 2
	Problem Space	<3	6	1	3	1	1	Rounds 2,4,5
	Answer	Correct	0 Correct 3 Wrong	3 Correct 0 Wrong	2 Correct 1 Wrong	3 Correct 0 Wrong	3 Correct 0 Wrong	Rounds 2,4,5
<b>Task 5</b>	Time (min)	<5	TNC	12	18	13	12	Rounds 2,5
	Erroneous Action	<5	22	6	9	6	4	Round 5
	Problem Space	<3	9	4	2	2	1	Round 5
	Answer	Correct	1 Correct 2 Wrong	3 Correct 0 Wrong	2 Correct 1 Wrong	3 Correct 0 Wrong	2 Correct 1 Wrong	Rounds 2,4

TNC ~ Task Not Completed

## APPENDIX J: CHA Survey

### Background Information

1. Have you personally done at least one community health assessment in the last 3 years?
  2. If 'No', have you utilized data from a community health assessment for your own program planning, policy development, and research in the last 3 years?
  3. Have you commissioned a community health assessment in the past 3 years?
  4. What best describes your organization?
- 

### Community Health Assessment Problem Solving

Your need to analyze health and population data sets in order to perform community health assessment analysis and answer the following 5 tasks below. The (electronic-format) data sets and variables available to you are:

#### Data Sets

Cancer Data Set	
Variable	Description
Age_ID	< 1, 1, 2, 3, 4, 5,.....85+
Geography_ID	Municipality
Sex_ID	Male, Female
Year_ID	1996, 1997, 1998, 1999, 2000
Disease_ID	ICD9-CM Disease Name
Numerical Measure	Number of Cancer Diagnoses

Inpatient Hospitalization Data Set	
Variable	Description
Age_ID	< 1, 1, 2, 3, 4, 5,.....85+
Geography_ID	Municipality
Sex_ID	Male, Female
Year_ID	1996, 1997, 1998, 1999, 2000
Disease_ID	ICD9-CM Disease Name
Numerical Measure	Number of Inpatient Visits

Outpatient Data Set		Death Data Set	
Variable	Description	Variable	Description
Age_ID 📌	< 1, 1, 2, 3, 4, 5,.....85+	Age_ID 📌	< 1, 1, 2, 3, 4, 5,.....85+
Geography_ID 📌	Municipality	Geography_ID 📌	Municipality
Sex_ID 📌	Male, Female	Sex_ID 📌	Male, Female
Year_ID 📌	1996, 1997, 1998, 1999, 2000	Year_ID 📌	1996, 1997, 1998, 1999, 2000
Disease_ID 📌	ICD9-CM Disease Name	Disease_ID 📌	ICD9-CM Disease Name
Numerical Measure	Number of Outpatient Visits	Numerical Measure	Number of Deaths


  

Population Data Set	
Variable	Description
Age_ID 📌	< 1, 1, 2, 3, 4, 5,.....85+
Geography_ID 📌	Municipality
Sex_ID 📌	Male, Female
Year_ID 📌	1996, 1997, 1998, 1999, 2000
Numerical Measure	Population

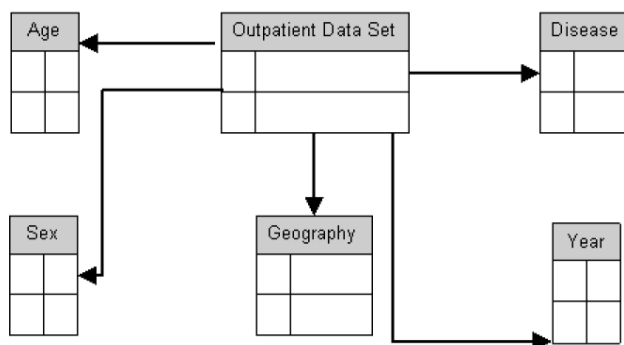
In addition to these data sets, you also have tables for some of the variables. The tables allow you to explore different levels of granularity. For example, given an instance of a Geography\_ID such as Aspinwall borough, the Geography table allows you to view data for its county, Allegheny.

### Variables

<i>Age</i>		<i>Geography</i>	
Age_ID 📌	Age_Bracket	Geography_ID 📌	County
1	1-4	Abbottstown borough	Adams
2	1-4	Arendtsville borough	Adams
3	1-4	Aleppo township	Allegheny
4	1-4	Aspinwall borough	Allegheny
5	5-14	Avalon borough	Allegheny
...	...	...	...

Disease		
Disease_ID 	Disease_Group	Disease_Category
Thyroiditis	Disorders of Thyroid Gland	Endocrine Disorders
Neoplasm of Colon	Neoplasm of Digestive Organs	Neoplasm
Intracerebral Hemorrhage	Cerebrovascular Disease	Circulatory System Disease
Acute Appendicitis	Appendicitis	Digestive System Disease
Appendicitis, unqualified	Appendicitis	Digestive System Disease
...	...	...

With these tables, the relationship between the outpatient data set (for example) and the variable tables could look as such:



You first combine the health data sets with the population data set, and calculate these additional (age-adjusted) numerical measures:

**cancer/100,000; inpatient visits/1,000; outpatient visits/1,000; deaths/100,000**

Below are 5 example community health assessment tasks. Using the data sets you have, please think of what information technologies (IT) you would use to solve community health assessment questions. For each task, please provide the specific tool(s) you would use (Please be as specific as possible by describing application name, ex. *SPSS Statistical Software* 12.0, instead of just *statistical software*) and how you would use these tools to solve the problem. Tools may include general purpose software applications or specific-purpose (full-feature) decision support technology. Your IT may run locally on your computer or remotely as a Web tool.

If you would not use IT software for a particular aspect of the task, please describe the manual process you would use.

## \*\*\*\*\*Part 1\*\*\*\*\*

### Task 1

*"How does the deaths/100,000 of Allegheny County in 1996 compare to the deaths/100,000 of each of the counties that border it?"*



### Task 2

*For this task Northwestern PA is defined by the following counties: Crawford, Erie, Forest, Mercer, Venango, Warren. Northeastern PA is defined by: Monroe, Pike, Susquehanna, Wayne, and Wyoming.*

*"What is the difference in cancer/100,000 of female "Malignant Neoplasm of Pancreas" in 1998 between northwestern and northeastern PA?"*



### Task 3

*"How does the inpatient/1,000 in 2000 of "complications occurring mainly in the course of labor and delivery" in Forest County compare to the counties that border it? For the county with the highest rate, what are the top 5 municipalities? Do all these municipalities border each other?"*



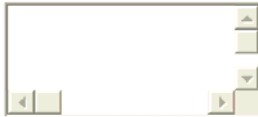
### Task 4

*"What are the top 5 "circulatory system" diagnoses of Inpatient/1,000 for Males Aged 65-74 in McKean County in 1999? For the diagnosis with the highest rate, how does this rate compare to the state-wide rate?"*



#### Task 5

*"What are the top 5 municipalities in Bucks County that have the highest deaths/100,000 from "Asthma" in 2000? How do these municipalities compare to the state-wide rate?"*




---

## \*\*\*\*\*Part 2\*\*\*\*\*

Now, for these same tasks, please provide the specific tool(s) you would use for each step (Please be as specific as possible by describing application name, ex. *SPSS Statistical Software 12.0*, instead of just *statistical software*). Tools may include software applications or full-feature decision support technology. If you would not use IT software for a step, please describe the manual process you would use.

*Example:*

<u>Step</u>	<u>Tool(s) Used</u>
<i>Identify bordering counties</i>	<i>ArcView GIS 3.3</i>

#### Task 1

*"How does the deaths/100,000 of Allegheny County in 1996 compare to the*

*deaths/100,000 of each of the counties that border it?"*

<u>Step</u>	<u>Tool(s) Used</u>
Access County level data	<input type="text"/>
Find deaths/100,000 in 1996	<input type="text"/>
Identify bordering counties	<input type="text"/>
Compare Border Counties to Allegheny County	<input type="text"/>

## Task 2

*For this task Northwestern PA is defined by the following counties: Crawford, Erie, Forest, Mercer, Venango, Warren. Northeastern PA is defined by: Monroe, Pike, Susquehanna, Wayne, and Wyoming.*

*"What is the difference in cancer/100,000 of female "Malignant Neoplasm of Pancreas" in 1998 between northwestern and northeastern PA?"*

<u>Step</u>	<u>Tool(s) Used</u>
Access County level data	<input type="text"/>
Create a custom set containing Northwestern PA	<input type="text"/>
Create a custom set containing Northeastern PA	<input type="text"/>
Find cancer/100,000 for 'Malignant Neoplasm of Pancreas' for females in 1998	<input type="text"/>
Compare Northwestern PA to Northeastern PA	<input type="text"/>

## Task 3

*"How does the inpatient/1,000 in 2000 of "complications occurring mainly in the course of labor and delivery" in Forest County compare to the counties that border it? For the county with the highest rate, what are the top 5 municipalities? Do all these municipalities border each other?"*

<u>Step</u>	<u>Tool(s) Used</u>
Access County level data	<input type="text"/>

Access Disease_Group level data	<input type="text"/>
Find inpatient/1,000 in 2000 for 'Complications occurring mainly in the course of labor and delivery'	<input type="text"/>
Identify bordering counties	<input type="text"/>
Compare Forest to bordering counties	<input type="text"/>
Find Top 5 municipalities of county with highest rate	<input type="text"/>
Determine if municipalities border one another	<input type="text"/>

#### Task 4

*"What are the top 5 "circulatory system" diagnoses of Inpatient/1,000 for Males Aged 65-74 in McKean County in 1999? For the diagnosis with the highest rate, how does this rate compare to the state-wide rate?"*

<u>Step</u>	<u>Tool(s) Used</u>
Access County level data	<input type="text"/>
Access Disease_Category level data	<input type="text"/>
Find inpatient/1,000 in 1999 for males aged 65-74 for 'circulatory system' disease for McKean County	<input type="text"/>
Compare highest rate to State-wide rate	<input type="text"/>

#### Task 5

*"What are the top 5 municipalities in Bucks County that have the highest deaths/100,000 from "Asthma" in 2000? How do these municipalities compare to the state-wide rate?"*

<u>Step</u>	<u>Tool(s) Used</u>
Access County level data	<input type="text"/>
Find deaths/100,000 in 2000 for 'asthma' in Bucks county	<input type="text"/>

Find Top 5 municipalities

Compare municipalities versus  
state-wide rate

---

Finish

Reset

## APPENDIX K: Recruitment Email for CHA Survey

Email Cover Letter

\*\*\*\*\*

Dear X,

I need your valuable help. I am looking for individuals who are familiar with the types of information technology (IT) used during community health assessments. I have created an online survey to help me identify these current information technologies (IT).

The link below will bring you to a web page that describes the survey.

I am looking for community health assessment researchers to complete the survey. This might be you if you have:

- \* Personally done at least one community health assessment in the last 3 years, OR
- \* Utilized data from a community health assessment for your own program planning, policy development, and research in the last 3 years, OR
- \* Commissioned a community health assessment in the past 3 years.

Please feel free to complete the survey if any of this applies to you. Also, if you can think of any additional people, please forward them this email. If this applies to you and you decide to complete the survey, I will pay you \$5 for completing the survey. The survey itself will take about 20 minutes.

The introduction page for the survey is:

<http://www.sovat.pitt.edu/Introduction.htm?SurveyCode=12345>

Thank You.

Matthew Scotch, MA  
Doctoral Candidate  
Center for Biomedical Informatics  
University of Pittsburgh

## APPENDIX L: Results from CHA Survey

### Data Management/Data Access

Task No	Step No	Most Popular Tool
1	1	Statistical Software
2	1	Statistical Software, Web-based Interface
3	1	Statistical Software
3	2	Statistical Software
4	1	Statistical Software
4	2	Statistical Software
5	1	Statistical Software
<b>Most Popular Overall</b>		<b>Statistical Software</b>

### Data Navigation

Task No	Step No	Most Popular Tool
1	2	Statistical Software
2	4	Statistical Software
3	3	Statistical Software
4	3	Statistical Software
5	2	Statistical Software
<b>Most Popular Overall</b>		<b>Statistical Software</b>

### Geographic Comparison

Task No	Step No	Most Popular Tool
1	4	Statistical Software
2	5	Statistical Software
3	5	Statistical Software
4	4	Statistical Software
5	4	Statistical Software
<b>Most Popular Overall</b>		<b>Statistical Software</b>

### Spatial Boundaries

Task No	Step No	Most Popular Tool
1	3	GIS Software
3	4	GIS Software
3	7	GIS Software
<b>Most Popular Overall</b>		<b>GIS Software</b>

### Spatial Modeling

Task No	Step No	Most Popular Tool
2	2	Statistical Software
2	3	Statistical Software
Most Popular Overall		Statistical Software

### Ranking Analysis (Top 5)

Task No	Step No	Most Popular Tool
3	6	Statistical Software
5	3	Statistical Software
Most Popular Overall		Statistical Software

## APPENDIX M: Apriori Algorithm for CHA Survey Results

### Apriori Algorithm Showing Association Rules for the Individual Software Applications

	Instances	Support	Confidence	Consequent	Antecedent 1	Antecedent 2	Antecedent 3
1	12	44.40	66.70	SPSS	ArcGIS		
2	11	40.70	72.70	ArcGIS	SPSS		
3	11	40.70	63.60	Excel	Web		
4	13	48.10	53.80	Web	Excel		
5	12	44.40	50.00	Web	ArcGIS		
6	11	40.70	54.50	Excel	SPSS		
7	11	40.70	54.50	ArcGIS	Web		
8	6	22.20	66.70	SPSS	ArcGIS	Web	
9	5	18.50	80.00	ArcGIS	SPSS	Web	
10	8	29.60	50.00	Excel	Conventional Map		
11	8	29.60	50.00	Web	Conventional Map		
12	8	29.60	50.00	Web	ArcGIS	SPSS	
13	4	14.80	75.00	SPSS	Excel	ArcGIS	
14	4	14.80	75.00	Web	Excel	ArcGIS	
15	5	18.50	60.00	Excel	SPSS	Web	
16	6	22.20	50.00	ArcGIS	SAS		
17	6	22.20	50.00	Web	SAS		
18	6	22.20	50.00	Conventional Map	SAS		
19	6	22.20	50.00	ArcGIS	Excel	SPSS	
20	6	22.20	50.00	Excel	ArcGIS	Web	
21	6	22.20	50.00	Web	Excel	SPSS	
22	3	11.10	66.70	ArcGIS	Epi-Info		
23	3	11.10	66.70	Excel	SPSS	Conventional Map	
24	3	11.10	66.70	Conventional Map	Web	SAS	
25	3	11.10	66.70	Web	Conventional Map	SAS	
26	3	11.10	66.70	Web	Excel	ArcGIS	SPSS
27	3	11.10	66.70	SPSS	Excel	ArcGIS	Web
28	3	11.10	66.70	ArcGIS	Excel	SPSS	Web
29	4	14.80	50.00	Excel	Other		
30	4	14.80	50.00	SPSS	Excel	Conventional Map	
31	4	14.80	50.00	Web	Excel	Conventional Map	
32	4	14.80	50.00	Excel	Web	Conventional Map	
33	4	14.80	50.00	SAS	Web	Conventional Map	
34	4	14.80	50.00	Excel	ArcGIS	SPSS	Web

**APPENDIX N: Informed Consent Form for Evaluation Study**

Current Approval Date: October 21, 2005  
Modification Approval Date: January XX, 2006  
Renewal Date: October 20, 2006  
University of Pittsburgh  
Institutional Review Board  
IRB# 0510032

**CONSENT TO ACT AS A PARTICIPANT IN A RESEARCH STUDY**

**TITLE:** INFORMATION TECHNOLOGY FOR COMMUNITY HEALTH ASSESSMENT  
RESEARCH

**PRINCIPAL INVESTIGATOR:**

Matthew Scotch, M.A.  
Ph.D. Candidate  
Center for Biomedical Informatics  
University of Pittsburgh  
8084 Forbes Tower  
Telephone: 412-647-7306

**CO-INVESTIGATORS:**

Bambang Parmanto, Ph.D.  
Assistant Professor  
Center for Biomedical Informatics  
University of Pittsburgh  
8084 Forbes Tower  
Telephone: 412-383-6649

Valerie Monaco, Ph.D.  
Assistant Professor  
Center for Biomedical Informatics  
University of Pittsburgh  
8084 Forbes Tower  
Telephone: 412-647-3064

Ravi K. Sharma, Ph.D.  
Assistant Professor  
Department of Behavioral and Community Health Sciences  
Graduate School of Public Health  
University of Pittsburgh  
PUBHL 228  
Telephone: 412-624-3615

**SOURCE OF SUPPORT:** None

***What is the purpose of this study?***

You are being asked to participate in a research study in which we will examine the use of information technology during community health assessment problem solving.

***Who is being asked to take part in this research study?***

You are being asked to take part in this research study because you are a potential researcher in the field of community health assessment. 15-40 individuals will participate in the study. Males and females age 18 years or older are being recruited for this study.

***What will my participation in this research study involve?***

If you agree to participate in the study, you may be asked to perform the following procedures in two different sessions (if this is your first time, a follow-up appointment may be scheduled before you leave):

**Experimental Procedures**

6. You will be asked to complete a pre-study form related to the future use of the video captured during your participation in the study.
7. You will be given instructions and asked to complete 2 practice tasks using an information technology system.
8. If you feel you are ready to use the system, you will be given a series of 4 or 5 tasks to complete. The actions you take with the mouse and the keyboard will be recorded via screen capture software, and any words you say out loud will be captured via a microphone.
9. Following the completion of all the tasks, you will be given a survey about your overall experience using the system. If this is your final session, you will be given an additional questionnaire related to your computer background.
10. Each session should take 2.5 hours to complete.

***What are the possible risks, side effects, and discomforts of this research study?***

There is no risk of physical injury associated with your participation in the study. Since we don't collect your personal identifiable information, participation in the study is not expected to involve the possible risk that your information is known to other individuals, although there always exists a risk for breach of confidentiality. The risk is minimized by keeping your research information confidential.

***Will I be paid if I take part in this research study?***

You will be given between \$10 and \$50, depending on the amount of participation in the study. You will receive \$10 after your first session. If you participate in a second session, you will be paid \$40 for completing the second session. If you attend the second session but are unable to complete it, you will receive \$10. In addition, any parking fees related to your participation in this study will be paid for by the study.

***Who will know about my participation in this research study?***

All records related to your involvement in this research study will be stored via a privately password-protected file on one of the researcher's computers. Your identity on these records will be indicated by a unique identification number rather than by your name, and the information linking these numbers with your identity will be kept separate from the research records (also password protected). Only the researchers listed on the first page of this page of this form and their staff will have access to your research records. Your research records will be destroyed when such is approved by the sponsor of this study or, as per University policy, at 5 years following study completion, whichever should occur first.

Any information about you obtained from this research will be kept as confidential (private) as possible. You will not be identified by name in any publication of research results unless you sign a separate form giving your permission (release). In unusual cases, your research records may be released in response to an order from a court of law.

***Who will have access to identifiable information related to my participation in this research study?***

In addition to the investigators listed on the first page of this consent form and their research staff, the following individuals will or may have access to identifiable information related to your participation in this research study:

Authorized representatives of the University of Pittsburgh Research Conduct and Compliance Office may review your identifiable research information for the purpose of monitoring the appropriate conduct of this research study. In unusual cases, the investigators may be required to release identifiable information related to your participation in this research study in response to an order from a court of law. If the investigators learn that you or someone with whom you are involved is in serious danger or potential harm, they will need to inform, as required by Pennsylvania law, the appropriate agencies.

***Is my participation in this research study voluntary?***

Your participation in this research study is completed voluntary. You do not have to take part in this research study and, should you change your mind, you can withdraw from the study at any time. Your current and future care at the University of Pittsburgh and any other benefits for which you qualify will be the same whether you participate in this study or not.

***May I withdraw, at a future date, my consent for participation in this research study?***

*If I agree to participation this research study, can I be removed from the study without my consent?*

**VOLUNTARY CONSENT**

By signing this form, I agree to participate in this research study. A copy of this consent form will be given to me.

---

Date

CERTIFICATION of INFORMED CONSENT

I certify that I have explained the nature and purpose of this research study to the above-named individual(s), and I have discussed the potential benefits and possible risks of study participation. Any questions of the individual(s) have about this study have been answered, and we will always be available to address future questions as they arise.

---

Printed Name of Person Obtaining Consent Role in Research Study

---

Signature of Person Obtaining Consent

---

Date

## **APPENDIX O: Tasks for Evaluation Study**

### **5 Tasks Given During SPSS-GIS Session**

*How does the outpatient rate per 1,000 of Crawford County in 1997 compare to the outpatient rates per 1,000 in 1997 of the different counties that border it?*

*For this task the Southern PA community is defined by the following counties: Adams, Chester, Franklin, Lancaster, and York.*

*The Central PA community is defined by the following counties: Centre, Huntingdon, Juniata, Mifflin, and Perry.*

*Compare the cancer incidence rate per 100,000 of female “Malignant neoplasm of colon” in 2000 between Southern PA and Central PA. Which counties not included in these communities border both of these two communities?*

*How does the cancer incidence rate per 100,000 in 1999 of Males Aged 65-74 in Centre County compare to the cancer incidence rates per 100,000 (in 1999 of Males Aged 65-74) of the different counties that border it?*

*For the county with the highest rate, how does this rate compare with the state-wide rate for cancer incidence per 100,000 in 1999 of Males Aged 65-74?*

*What are the top 5 counties of deaths per 100,000 of “Circulatory system” diseases in 2000? Does one part of the state appear to contain the top 5 counties?*

*How does the Inpatient LOS (Length of Stay) per 1,000 in 2000 for females compare between McKean and Mercer Counties? For the county with the higher rate, what are its top 5 municipalities of Inpatient LOS per 1,000 in 2000 for females? Do all these municipalities border one another?*

### **5 Tasks Given During SOVAT Session**

*How does the outpatient rate per 1,000 of Warren County in 1998 compare to the outpatient rates per 1,000 in 1998 of the different counties that border it?*

*For this task the Eastern PA community is defined by the following counties: Bucks, Carbon, Lehigh, Monroe, and Northampton.*

*The Northern PA community is defined by the following counties: Bradford, McKean, Potter, Susquehanna, and Tioga.*

*Compare the cancer incidence rate per 100,000 of female “Malignant neoplasm of colon” in 2000 between Eastern PA and Northern PA. Which counties not included in these communities border both of these two communities?*

*How does the cancer incidence rate per 100,000 in 1999 of Males Aged 75-84 in Indiana County compare to the cancer incidence rates per 100,000 (in 1999 of Males Aged 75-84) of the different counties that border it?*

*For the county with the highest rate, how does this rate compare with the state-wide rate for cancer incidence per 100,000 in 1999 of Males Aged 75-84?*

*What are the top 5 counties of deaths per 100,000 of “Infectious and parasitic diseases” in 2000? Does one part of the state appear to contain the top 5 counties?*

*How does the Inpatient LOS (Length of Stay) per 1,000 in 2000 for females compare between Elk and Clarion Counties? For the county with the higher rate, what are its top 5 municipalities of Inpatient LOS per 1,000 in 2000 for females? Do all these municipalities border one another?*

Subject Recruitment Flier

# **CALLING ALL GSPH Students**

## **Are you proficient with SPSS?**

### **I Need Your Help in a Research Study to Evaluate Software**

You will be asked to use software to complete assigned tasks, and to fill out a background survey and one questionnaire. Each appointment may last 2.5 hours. Most likely, you will have two separate sessions. I will pay you a minimum of \$10 with the potential to earn up to \$50.

The whole experiment will be conducted on campus at the University of Pittsburgh.

Contact Matthew Scotch at (412) 647-7306 (email: [scotch@cbmi.pitt.edu](mailto:scotch@cbmi.pitt.edu)) for more information or for participation. Please leave your name and a phone number so you can be contacted for the first appointment.

## BIBLIOGRAPHY

1. Scotch M and Parmanto B, *Development of SOVAT: A Numerical-Spatial Decision Support System for Community Health Assessment Research*. International Journal of Medical Informatics, 2005. **In Press**.
2. Tufte E, *Visual Explanations*. 1997, Cheshire, CT: Graphics Press.
3. Gorry AG and Scott Morton MS, *A Framework for Management Information Systems*. Sloan Management Review, 1971. **13**(1): p. 55-70.
4. Sprague RH and W. HJ, *Bit by Bit: Toward Decision Support Systems*. California Management Review, 1979. **22**(1): p. 60-68.
5. Methlie LB, *Data management for decision support systems*. ACM SIGOA Newsletter (Selected papers on decision support systems from the 13th Hawaii International Conference on System Sciences), 1980. **1**(4-5): p. 40-46.
6. Gorry AG, *The development of Managerial Models*. Sloan Management Review, 1971. **12**(2): p. 1-16.
7. Wagacha PW, *Modeling and Decision Support*, U.o.N. School of Computing & Informatics, Editor. 2004.
8. Dutta A and Basu A, *An Artificial Intelligence Approach to Model Management in Decision Support Systems*. IEEE Computer, 1984. **17**(9): p. 89-97.
9. Keen PGW, *"Interactive" Computer Systems for Managers: A Modest Proposal*. Sloan Management Review, 1976. **18**(Fall): p. 1-17.
10. Aiken MW, Liu-Sheng OR, and Vogel DR, *Integrating Expert Systems with Group Decision Support Systems*. ACM Transactions on Information Systems, 1991. **9**(1): p. 75-95.
11. Al-Ani I, Cooley RE, and Awad EM. *From Decision Support to Expert Systems*. in *Proceedings of the conference on The 1987 ACM SIGBDP-SIGCPR Conference*. 1987. Coral Gables, Florida.
12. Angehrn AA and L. H-J., *Intelligent Decision Support Systems: A Visual Interactive Approach*. Interfaces, 1990. **20**(6): p. 17-28.
13. Chen MS, Chau CC, and Kabat WC. *Decision Support Systems: A Rule-Based Approach*. in *Proceedings of the 1985 ACM annual conference on The range of computing*. 1985.
14. Davis R, *A DSS for Diagnosis and Therapy*. Data Base, 1977. **8**(3): p. 58-72.
15. El-Najdawi MK and Stylianou AC, *Expert Support Systems: Integrating AI Technologies*. Communications of the ACM, 1993. **36**(12): p. 55-103.
16. Gorry AG and Krumland RB, *Artificial Intelligence Research and Decision Support Systems*, in *Building Decision Support Systems*, Bennett JL, Editor. 1983, Addison-Wesley: Reading, MA. p. 205-219.
17. Jarke M and Vassiliou Y, *Coupling Expert Systems with Database Management Systems*, in *Artificial Intelligence Applications for Business*, Reitman W, Editor. 1984, Ablex Publishing Corporation: Norwood, NJ. p. 65-85.
18. Jin B, Hurson AR, and Miller LL. *Neural Network-Based Decision Support for Incomplete Database Systems: Knowledge Acquisition and Performance Analysis*. in *Proceedings of the conference on Analysis of neural network applications*. 1991. Fairfax, Virginia.
19. Kimbrough SO, et al., *The Coast Guard's KSS Project*. Interfaces, 1990. **26**(6): p. 5-16.

20. Norrie DH, et al. *A knowledge-based decision support system for flexible manufacturing*. in *Proceedings of the second international conference on Industrial and engineering applications of artificial intelligence and expert systems*. 1989. Tullahoma, Tennessee.
21. Owens HD and Philippakis AS, *Inductive Consistency in knowledge-based decision support systems*. *Decision Support Systems*, 1995. **13**(2): p. 167-181.
22. Reitman W, *Applying Artificial Intelligence to Decision Support: Where Do Good Alternatives Come From?*, in *Decision Support Systems*, Ginzberg MJ, Reitman W, and Stohr EA, Editors. 1982, North-Holland Publishing Company: Amsterdam. p. 155-174.
23. Turban E and Watkins PR, *Integrating Expert Systems and Decision Support Systems*. *MIS Quarterly*, 1986. **10**(2): p. 121-136.
24. Barron T and Sharia AN, *Data Requirements in Statistical Decision Support Systems: Formulation and Some Results in Choosing Summaries*. *Decision Support Systems*, 1995. **15**(4): p. 375 - 388.
25. Centers for Disease Control and Prevention - Interactive Atlas of Reproductive Health, <http://www.cdc.gov/reproductivehealth/gisatlas/>. 2004.
26. Thomas AJ and Carlin BP, *Late detection of breast and colorectal cancer in Minnesota counties: an application of spatial smoothing and clustering*. *Statistics in Medicine*, 2003. **22**(1): p. 113-127.
27. Kimball, R., *A Dimensional Modeling Manifesto*. *DBMS Magazine*, 1997(August 1997).
28. Koutsoukis N-S, Mitra G, and Lucas C, *Adapting on-line analytical processing for decision modeling: the interaction of information and design technologies*. *Decision Support Systems*, 1999. **26**(1): p. 1-30.
29. Honeywell, *DOmain Modeling Environment*. 2004. p. <http://www.htc.honeywell.com/dome/>.
30. Keenan P, *Using a GIS as a DSS Generator*, in *Working Paper MIS*. 1997, University College Dublin. p. 95-99.
31. Carlson ED, et al. *The Design and Evaluation of an Interactive Geo-Data Analysis and Display System*. in *Proceedings of the IFIP Congress 1974*. 1974.
32. Lamb FC and Hough B. *GIS Use in America's War Against Terrorism*. in *Twenty-second Annual ESRI User Conference*. 2002. San Diego, California.
33. MapQuest, [www.mapquest.com](http://www.mapquest.com). 2004.
34. Krivoruchko K, Gotway CA, and Zhigimont A. *Statistical Tools for Regional Data Analysis Using GIS*. in *GIS '03*. 2003. New Orleans, LA.
35. Apte C, et al., *Evolving data mining into solutions for insights: Business applications of data mining*. *Communications of the ACM*, 2002. **45**(8): p. 49-53.
36. Estivill-Castro V and Lee I. *Fast spatial clustering with different metrics and in the presence of obstacles*. in *Proceedings of the ninth ACM international symposium on Advances in geographic information systems*. 2001. Atlanta, Georgia.
37. Gaines BR. *Organizational modelling and problem solving using object-oriented knowledge representation server and visual language*. in *Conference proceedings on Organizational computing systems*. 1991. Atlanta, Georgia.
38. Guo D, Peuquet D, and Gahegan M. *Data clustering: Opening the black box: interactive hierarchical clustering for multivariate spatial patterns*. in *Proceedings of the tenth ACM international symposium on Advances in geographic information systems*. 2002. McLean, Virginia.

39. Han J, Koperski K, and Stefanovic N. *GeoMiner: a system prototype for spatial data mining*. in *Proceedings of the 1997 ACM SIGMOD international conference on Management of data*. 1997. Tucson, Arizona.
40. Indulska M and Orlowska ME. *Data clustering: Gravity based spatial clustering*. in *Proceedings of the tenth ACM international symposium on Advances in geographic information systems*. 2002. McLean, VA.
41. Miller H and Han J, *Geographic Data Mining and Knowledge Discovery*. 2001, London, England: Taylor & Francis.
42. Qian Y, Zhang G, and Zhang K. *Data mining: FAÇADE: a fast and effective approach to the discovery of dense clusters in noisy spatial data*. in *Proceedings of the 2004 ACM SIGMOD international conference on Management of data*. 2004. Paris, France.
43. Qian Y and Zhang G. *Data mining (DM): GraphZip: a fast and automatic compression method for spatial data clustering*. in *Proceedings of the 2004 ACM symposium on Applied computing*. 2004. Nicosia, Cyprus.
44. Rosales R, Achan K, and Frey B. *Learning to cluster using local neighborhood structure*. in *Twenty-first international conference on Machine learning*. 2004. Banff, Alberta, Canada.
45. Son E-J, et al. *A spatial data mining method by clustering analysis*. in *Proceedings of the sixth ACM international symposium on Advances in geographic information systems*. 1998. Washington, D.C.
46. Hoebe CJP, et al., *Space-time cluster analysis of invasive meningococcal disease*. *Emerg Infect Dis*, 2004.
47. Han J, Kamber M, and Tung AKH, *Spatial clustering methods in data mining: A survey*, in *Geographic Data Mining and Knowledge Discovery*, Miller HJ and Han J, Editors. 2001, Taylor & Francis: New York.
48. Mahinthakumar G, et al. *Multivariate geographic clustering in a metacomputing environment using Globus*. in *Proceedings of the 1999 ACM/IEEE conference on Supercomputing*. 1999. Portland, Oregon.
49. Gao S, Paynter J, and Sundaram D. *Flexible Support for Spatial Decision-Making*. in *Proceedings of the 37th Hawaii International Conference on System Sciences*. 2004.
50. Bedard, Y., et al., *Integrating GIS components with knowledge discovery technology for environmental health decision support*. *Int J Med Inf*, 2003. **70**(1): p. 79-94.
51. Bedard Y, *JMAP Spatial OLAP: Innovative technology to support intuitive and interactive exploration and analysis of spatio-temporal multidimensional data*, K. Technologies, Editor. 2005: Montreal, PQ. p. 1-14.
52. Tufte, E.R., *Visual and Statistical Thinking: Displays of Evidence for Making Decisions*, in *Envisioning Information*. 1990, Graphics Press: Cheshire, CT.
53. Richard JB, et al. *A Web-Based GIS for Health Care Decision Support*. in *American Medical Informatics Association 2005 Symposium*. 2005. Washington, DC.
54. Bapna S and Gangopadhyay A, *A Web-Based GIS for Analyzing Commercial Motor Vehicle Crashes*. *Information Resources Management Journal*, 2005. **18**(3): p. 1-12.
55. Hoskins RE, et al., *EpiQMS: An Internet Application for Access to Public Health Data for Citizens, Providers, and Public Health Investigators*. *Journal of Public Health Management Practice*, 2002. **8**(3): p. 30-36.

56. Hernandez V, Gohring W, and Hopmann C, *Sustainable Decision Support for Environmental Problems in Developing Countries: Applying Multi-Criteria Spatial Analysis on the Nicaragua Development Gateway in niDG*, in EU-LAT. 2004.
57. May M and Savinov A. *SPIN! - An Enterprise Architecture for Spatial Data Mining*. in KES. 2003.
58. Nehme CC and Simoes M. *Spatial Decision Support System for Land Assessment*. in *Proceedings of the seventh ACM international symposium on Advances in geographic information systems*. 1999. Kansas City, MO.
59. Goddard S, et al., *Geospatial Decision Support for Drought Risk Management*. Communications of the ACM, 2003. **46**(1): p. 35-37.
60. HRSA Geospatial Data Warehouse, <http://datawarehouse.hrsa.gov/>. 2004: Rockville, MD.
61. National Cancer Institute GIS Database Project, <http://gis.cancer.gov/nci/database.html>.
62. Kulldorff M and et al, *Breast Cancer Clusters in Northeast United States: A Geographic Analysis*. American Journal of Epidemiology, 1997. **146**: p. 161-170.
63. Kulldorff M and et al, *SaTScan v 2.1, Software for the Spatial and Space-Time Scan Statistics*. 1998, National Cancer Institute: Bethesda, MD.
64. Visual OLAP, <http://www.visualolap.com/features.asp?id=6>.
65. Spofford G, *MDX Solutions: With Microsoft SQL Server Analysis Services*. 2001: John Wiley & Sons.
66. Gorla N, *Features to Consider in a Data Warehousing System*. Communications of the ACM, 2003. **46**(11): p. 111-115.
67. Kimball R, *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses*. 1996: John Wiley & Sons.
68. Lewis JR, *IBM Computer Usability Satisfaction Questionnaires: Psychometric Evaluation and Instructions for Use*. International Journal of Human-Computer Interaction, 1995. **7**(1): p. 57-78.
69. Davis F, *Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology*. MIS Quarterly, 1989. **13**(3): p. 319-340.
70. Mullet K and Sano D, *Designing Visual Interfaces: Communication Oriented Techniques*. 1995, Mountain View, CA: SunSoft Press. 273.
71. Sharma RK, *Putting the Community Back in Community Health Assessment: A Process and Outcome Approach with a Review of Some Major Issues for Public Health Professionals*. Journal of Health & Social Policy, 2003. **16**(3): p. 19-33.
72. Sadan E and Churchman A, *Process-Focused and Product-Focused Community Planning: Two Variations of Empowering Professional Practice*. Community Development Journal, 1997. **32**(1): p. 3-16.
73. Institute of Medicine, *Improving Health in the Community: A Role for Performance Modeling*. Committee on Using Performance Monitoring to Improve Community Health, ed. Durch JS, Bailey LA, and Stoto MA. 1997, Washington, DC: National Academy Press.
74. Friedman DJ, et al., *Assessment Partnerships Between Managed Care and Public Health: The Massachusetts Experience*. Journal of Public Health Management and Practice, 2002. **8**(4): p. 77-84.
75. Spradley BW, *Community Health Nursing: Concepts and Practice*. 1990, Glenview, IL: Scott, Foresman and Company.

76. Birmingham Foundation, *Birmingham Foundation*. 2005: Pittsburgh, PA.
77. Birmingham Foundation, *Reweaving the Social Fabric — A Community Health Assessment of South Pittsburgh*. 2005: Pittsburgh, PA.
78. Curtis DC, *Community Health Assessment in Kansas*. Journal of Public Health Management and Practice, 2002. **8**(4): p. 20-25.
79. Plescia M, Koontz S, and Laurent S, *Community assessment in a vertically integrated health care system*. American Journal of Public Health, 2001. **91**(5): p. 811-814.
80. Caley LM, *Using Geographic Information Systems to Design Population-Based Interventions*. Public Health Nursing, 2004. **21**(6): p. 547-554.
81. Cromley EK and McLafferty SL, *GIS and Public Health*. 2002, New York: The Guilford Press.
82. Mowat D, et al. *Improving health surveillance in Canada - What are the needs?* in *Information Technology and Community Health (ITCH)*. 2000. Victoria, BC, Canada.
83. Turning Point Initiative. 2005, <http://www.turningpointprogram.org/>.
84. Turning Point Initiative, *Information Technology Collaborative*. 2005, <http://www.turningpointprogram.org/Pages/infotech.html>.
85. Magruder C, et al., *Using Information Technology to Improve the Public Health System*. Journal of Public Health Management & Practice, 2005. **11**(2): p. 123-130.
86. Washington State Department of Health, *Information Technology: Reliable Information for Better Health Programs*. 2002: <http://www.doh.wa.gov/PHIP/documents/PHIP2002/2002PHIP8.pdf>.
87. Washington State Department of Health, *Public Health Information Technology Committee Assessment*. 2005: <http://www.doh.wa.gov/phip/InfoTech/Assess.htm>.
88. Westrum F, *PHIT Assessment 2002*, Scotch M, Editor. 2005: Tumwater, WA.
89. National Association of County and City Health Officials, *National Profile of Local Public Health Agencies*. 2005: A project supported through a cooperative agreement (number 302718) between the National Association of County and City Health Officials and the Centers for Disease Control and Prevention.
90. Centers for Disease Control, *National Public Health Performance Standards Program (NPHPSP)*. 2005, <http://www.phppo.cdc.gov/nphpsp/index.asp>: Atlanta, GA.
91. Corso L, *NPHPSP Survey*, Scotch M, Editor. 2005: Atlanta, GA.
92. Centers for Disease Control, *NPHPSP State/Local Public Health System Performance Assessment Instruments: Overall National Report for Responding Local Jurisdictions and States Summary of Performance Scores by EPHS and Indicators*. 2005.
93. Corso L, *NPHPSP Survey - GIS responses*, Scotch M, Editor. 2005: Atlanta, GA.
94. Morgan GA, *Sampling and External Validity*. Journal of the American Academy of Child & Adolescent Psychiatry, 1999. **38**(8): p. 1051-1053.
95. Braithwaite D, et al., *Using the Internet to conduct surveys of health professionals: a valid alternative?* Family Practice, 2003. **20**(5): p. 545-551.
96. Chung K, Yang DH, and Bell R, *Health and GIS: Toward Spatial Statistical Analyses*. Journal of Medical Systems, 2004. **28**(4): p. 349-360.
97. Agrawal R and Srikant R. *Fast Algorithms for Mining Association Rules*. in *Proc. 20th Int. Conf. Very Large Data Bases (VLDB)*. 1994: Morgan Kaufmann.
98. Bunker E. *Beyond the Spreadsheet: Usability of a Prototype Interface for Accessing and Visualizing Public Health Data*. in *Presentation at the National Library of Medicine Director's Meeting*. 2005. Bethesda, MD.

99. Hart M and Porter G, *The Impact of Cognitive and Other Factors on the Perceived Usefulness of OLAP*. Journal of Computer Information Systems, 2004. **45**(1): p. 47-56.
100. Davis F, *A Technology Acceptance Model for Empirically Testing New End-User Information Systems: Theory and Results*, in Sloan School of Management. 1986, MIT.
101. Zeng X, *Evaluation and Enhancement of Web Content Accessibility for Persons with Disabilities*, in Health Information Management. 2004, University of Pittsburgh: Pittsburgh, PA. p. 173.
102. Diaz-Uriarte R, *The analysis of cross-over trials in animal behavior experiment: review and guide to the statistical literature*. 2001, Madrid, Spain: Ramón Díaz-Uriarte.
103. Dallal GE, *The Computer-Aided Analysis of Crossover Studies*. 2000.
104. Vorosmarty CJ, et al., *Geospatial indicators of emerging water stress: an application to Africa*. Ambio, 2005. **34**(3): p. 230-6.
105. Briggs D, *The role of GIS: coping with space (and time) in air pollution exposure assessment*. Journal of Toxicology & Environmental Health Part A, 2005. **68**(13-14): p. 1243-61.
106. Jenks RH and Malecki JM, *GIS--a proven tool for public health analysis*. Journal of Environmental Health, 2004. **67**(3): p. 32-4.
107. Kaminska IA, Oldak A, and Turski WA, *Geographical Information System (GIS) as a tool for monitoring and analysing pesticide pollution and its impact on public health*. Annals of Agricultural & Environmental Medicine, 2004. **11**(2): p. 181-4.
108. Miller C, *The Use of a GIS to Compare the Land Areas Captured by Very Basic and Complex Wellhead Protection Area Models*. Journal of Environmental Health, 2005. **68**(4): p. 21-6.
109. Taylor B, et al., *Proximity to Pollution Sources and Risk of Amphibian Limb Malformation*. Environmental Health Perspectives, 2005. **113**(11): p. 1497-1501.
110. Waring S, et al., *The utility of geographic information systems (GIS) in rapid epidemiological assessments following weather-related disasters: methodological issues based on the Tropical Storm Allison Experience*. International Journal of Hygiene & Environmental Health, 2005. **208**(1-2): p. 109-16.