

# Final Report

for the SBIR Phase I research effort entitled:

*Spatial Data Mining Toolkit for Generating MSDS (aka TopoAssistant)*  
(Topic No. A03-129)

June 9, 2004

Under

**Contract No. W9132V-04-C-0015**

Issued and Administered by:  
US Army Topographic Engineering Center  
7701 Telegraph Road  
Alexandria, VA 22315-3864

Contractor: Architecture Technology Corporation  
9971 Valley View Road  
Eden Prairie, MN 55344  
<http://www.atcorp.com>

Principal Investigator: Sid Kudige  
Phone No.: (952) 829-5864 (extn. 163)  
E-mail: [skudige@atcorp.com](mailto:skudige@atcorp.com)

Effective Date of Contract: December 17, 2003  
Length of Contract: December 17, 2003 to June 16, 2004.

# TABLE OF CONTENTS

<b>1.</b>	<b>INTRODUCTION .....</b>	<b>1</b>
1.1	IDENTIFICATION AND SIGNIFICANCE OF THE PROBLEM .....	1
1.2	PHASE I TECHNICAL OBJECTIVES .....	2
1.3	TOPOASSISTANT INNOVATIONS .....	2
1.4	SUMMARY OF THE PHASE I RESULTS .....	6
1.5	REPORT OUTLINE .....	6
<b>2.</b>	<b>PHASE I RESULTS .....</b>	<b>6</b>
2.1	PHASE I PROTOTYPE IMPLEMENTATION APPROACH .....	6
2.2	RESULTS OBTAINED BY APPLYING PHASE I PROTOTYPE TO TEC DATASET .....	8
2.3	PHASE I PROTOTYPE SCALABILITY AND PERFORMANCE .....	18
2.4	TECHNICAL CHALLENGES IDENTIFIED DURING PHASE I .....	18
2.4.1	<i>Flexible and Compatible TopoAssistant Implementation</i> .....	18
2.4.2	<i>Scalability Challenge</i> .....	18
2.4.3	<i>Modeling Spatial Patterns</i> .....	19
2.4.4	<i>Automation of spatial data mining methods</i> .....	21
2.4.5	<i>MSDS Verification/Error Detection and Data Enhancement Agent Synthesis</i> .....	22
2.4.6	<i>Merging Spatial Datasets from Disparate Sources</i> .....	23
2.4.7	<i>Provide Actionable Information to the Topographer</i> .....	23
2.5	TECHNICAL APPROACH TO ADDRESS THE TECHNICAL CHALLENGES .....	23
2.5.1	<i>Flexible TopoAssistant Implementation</i> .....	24
2.5.2	<i>Designing SDM Level Algorithms, Geometry Engine, Database Level and System Level Optimizations</i> .....	25
2.5.3	<i>Designing Verification/Error Detection and Data Enhancement/Tailoring Agents</i> .....	28
2.5.4	<i>Designing the Topographer's GUI</i> .....	32
<b>3.</b>	<b>RELATED WORK .....</b>	<b>33</b>
<b>4.</b>	<b>COMMERCIALIZATION STRATEGY .....</b>	<b>35</b>
4.1	COMMERCIALIZATION STRATEGY .....	35
4.2	COMPANY BACKGROUND .....	36
<b>5.</b>	<b>CONCLUSION .....</b>	<b>37</b>
<b>6.</b>	<b>REFERENCES .....</b>	<b>38</b>

## **PREFACE**

The final report describes in detail the work that has been performed by Architecture Technology Corporation (ATC) during the base period of this Phase I SBIR Topic No. A03-129 entitled *Spatial Data Mining Toolkit for Generating MSDS aka TopoAssistant*. This document was prepared by the Principal Investigator (PI) of this effort, Sid Kudige.

TEC's contract number and the COR are listed below.

### **Contract No. W9132V-04-C-0015**

Issued and Administered by:  
US Army Topographic Engineering Center  
7701 Telegraph Road  
Alexandria, VA 22315-3864

### **Contracting Officers Representative (COR)**

CEERD-TR-R, Dr James A Shine  
7701 Telegraph Road  
Alexandria, VA 22315-3864

## **SUMMARY (ABSTRACT)**

The geospatial information requirements for supporting Army applications such as battlefield situational awareness, tactical decision aids, and map backgrounds may not be adequately specified until close to the commencement of the mission. Hence, rapid and timely production of Mission Specific Data Sets (MSDS) is imperative to be responsive to mission needs. There is a need therefore for an automated tool to support rapid generation of MSDS with the required degree of fidelity from lower fidelity data. Spatial data mining techniques that have been developed in the recent past represent a key enabling technology for building tools to assist a topographer in rapidly generating MSDS feature data. Leveraging a number of novel spatial data mining techniques, Architecture Technology Corporation will develop an automation tool called TopoAssistant (Topographer's Assistant) to assist Army topographers in building "just-in-time" MSDS.

The Phase I SBIR effort successfully established the technical feasibility of the TopoAssistant approach by building a limited prototype of the tool and demonstrating its capabilities for map error detection and feature attribution. Building upon the results of the Phase I effort, the Phase II effort will implement a full-scale operational prototype of the TopoAssistant tool by integrating spatial data mining techniques with commercially available GIS software.

The TopoAssistant tool developed by this SBIR effort will significantly reduce the time and effort expended by Army topographers in generating MSDS to support operational mission requirements. It will enable them to address the responsiveness requirements for the Objective Force. This tool directly addresses the needs for applications such as Intelligence Preparation for the Battlefield (IPB), tactical decision aids (TDAs), corridor analysis, and route planning within the Future Combat Systems. National Geospatial Intelligence Agency (NGA) topographers also can use the TopoAssistant tool to generate higher fidelity maps. The US Census Bureau can use the TopoAssistant software for error detection/feature attribution of their maps. In the civilian sector, mapping companies such as Mapquest and Mapblast can increase accuracy of their maps and reduce positional and topological errors by using the TopoAssistant tool.

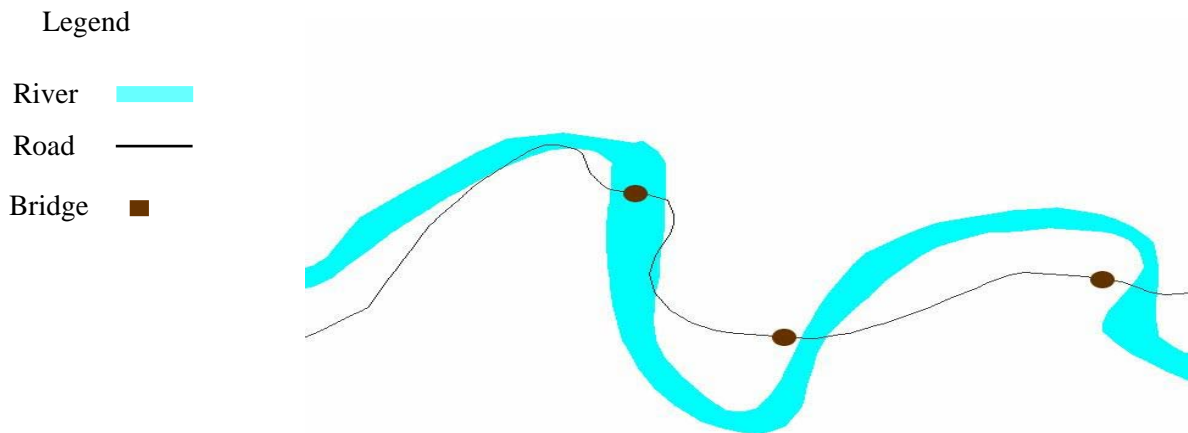
# 1. Introduction

## 1.1 Identification and Significance of the Problem

Maps are crucial for supporting battlefield situational awareness, routing and navigation functionality and precision targeting of enemy assets etc. In the current Iraq war precision targeting with help from accurate and high fidelity maps increased the lethality of Army weapons and reduced casualties. However, a large amount of expensive and non-repeatable manual effort (up to 2400 hours) is required to prepare accurate, up-to-date and high-fidelity Mission Specific Data Sets (MSDS) from lower fidelity data sources [Kabinier TEC].

Because of the labor cost involved in preparing MSDS and the limited time available, MSDS often contains errors and missing or out of date features. To address these concerns, the Topographic Engineering Center (TEC) has identified the need for effective software tools to reduce the manual labor needed for producing accurate and up-to-date MSDS. To minimize the manual effort and significantly automate routine tasks of MSDS preparation, Architecture Technology Corporation (ATC) proposes to develop a software tool called **TopoAssistant** (Topographer's Assistant) to assist Army topographers in rapid and automated refinement/production of high fidelity MSDS. The TopoAssistant will leverage groundbreaking **Spatial Data Mining** techniques (Shekhar et al.2004) developed by ATC researchers and Prof. Shashi Shekhar (consultant on this effort) at the University of Minnesota.

The Phase I component of this SBIR effort successfully built a concept demonstration prototype of TopoAssistant to establish the technical feasibility of the TopoAssistant tool for significantly automating the MSDS refinement process.



**Figure 1: Zoomed in View of Errors Detected in TEC's Korea Dataset**

Figure 1 depicts map errors detected automatically by ATC's Phase I TopoAssistant prototype in a Korea dataset provided to us by TEC personnel. These map errors cannot be detected manually, as they are not visible to the naked eye without considerable zooming. The figure shows a subset of roads, rivers and bridges. It is interesting to note that the location of some bridges does not coincide with intersection points of roads and rivers. This may indicate positional error in the locations of bridges, roads or rivers. In addition roads cross rivers in some locations with no

bridges in the vicinity. These spatial patterns indicate positional error or road/river and/or missing bridge features on the map. If left undetected, map errors of this kind could cause havoc with various automated systems and tools which use this map information to perform functions like precision targeting of bridges, routing and navigation etc.

It is instructive to examine the generic steps associated with producing the MSDS feature data in greater detail to underscore the limitations of the traditional manual process and to highlight the need for an automated tool like TopoAssistant to support MSDS production. The following paragraph describes the steps in the MSDS generation process.

In the **requirements analysis** phase, a specification detailing such properties as the geographic extents, resolution, feature and attribute set for the MSDS is produced. In the **data acquisition and import** phase, the lower fidelity source data, e.g. Feature Foundation Data (FFD) is selected and imported into the MSDS generation environment. In the **data enhancement and tailoring phase**, feature density is intensified (more instances of bridge, road etc are added) and missing attribute values are added (missing soil type, etc). Feature density could also be generalized or thinned in this phase. The next phase is the **verification/deconfliction** phase during which positional, topological and other errors within features like roads and rivers are rectified. After all the data deficiencies have been addressed, the data set is **exported** from the MSDS refinement and generation systems environment in a format appropriate for mission applications using the data.

Data Enhancement and Tailoring, and Verification/Error Detection are two of the most complex and time consuming steps in the manual MSDS generation process employed today. These steps currently rely on human reasoning resulting in expensive, non-repeatable MSDS production efforts as evidenced by data presented by Debra Kabinier of the US Army [Kabinier]. Spatial data mining techniques [Shekhar et al. 2004] that have been developed in the recent past represent a key enabling technology for building tools to assist a topographer in rapidly generating MSDS feature data. ATC will leverage spatial data mining techniques and integrate them with open source and commercially available GIS product offerings to produce the TopoAssistant tool for assisting Army topographers during the MSDS refinement process.

## 1.2 Phase I Technical Objectives

The Phase I SBIR effort established the technical feasibility of using spatial data mining techniques to build the TopoAssistant software tool to assist topographers and significantly automate the MSDS refinement process. The following were the specific technical objectives of the Phase I effort.

1. Requirements-driven selection of the spatial data mining techniques that need to be incorporated within TopoAssistant to meet the Army TEC's needs for MSDS refinement.
2. Design of the software architecture for TopoAssistant.
3. Concept demonstration of TopoAssistant through a rapid prototype to establish implementation feasibility of the toolkit.
4. Detailed design of the TopoAssistant software, including the development of a design specification document for use during the Phase II build of the toolkit (option contract).

## 1.3 TopoAssistant Innovations

Underlying our approach for building the TopoAssistant tool are two major innovations:

- A set of novel **spatial data mining techniques** including those developed by Prof. Shashi Shekhar at the University of Minnesota.
- **Open extensible TopoAssistant architecture** which allows the use of open source public domain software, COTS components and common standards (OGIS, C/JMTK)

### **Spatial Data Mining Techniques**

In the following paragraphs, we initially elaborate on the need for spatial data mining techniques by describing why classical data mining techniques are not adequate in the spatial domain. We then describe the various spatial data mining techniques that will be used to build a full-scale operational prototype of the TopoAssistant tool.

Differences between classical and spatial data mining are similar to the difference between classical and spatial statistics. First, spatial data is embedded in a continuous space, whereas classical data sets are often discrete. Second, spatial patterns are often local whereas classical data mining techniques often focus on global patterns. Finally, one of the common assumptions in classical statistical analysis is that data samples are independently generated. When it comes to the analysis of spatial data, however, the assumption about the independence of samples is generally false because spatial data tends to be highly autocorrelated. For example, people with similar characteristics, occupation and background tend to cluster together in the same neighborhoods. In spatial statistics this tendency is called spatial autocorrelation. Ignoring spatial autocorrelation when analyzing data with spatial characteristics may produce hypotheses or models that are inaccurate or inconsistent with the data set. Thus classical data mining algorithms often perform poorly when applied to spatial data sets. New methods are needed to analyze spatial data to detect spatial patterns.

We have identified four major classes of promising spatial data mining techniques to use as building blocks for TopoAssistant. The following paragraphs list and summarize the four classes of spatial data mining techniques of interest, i.e., Classification/Location Prediction, Spatial Outlier Detection, Spatial Co-Location, and Spatial Clustering. Further details are presented in [Shekhar et al 2004].

**Classification/Location Prediction:** This set of spatial data mining techniques is concerned with the discovery of a model (or a set of rules) to infer either the location or the classification of a spatial phenomenon from the maps of other spatial features. For instance, given a grid on a map these techniques derive rules for predicting a missing feature such as a bridge, in that grid. This is location prediction. Alternatively, given a feature located within a specified grid these techniques derive rules for predicting a missing attribute of the feature, e.g., soil type. This is class (or attribute) prediction.

**Spatial Outlier Detection:** Outliers have been informally defined as observations in a data set which appear to be inconsistent with the remainder of that set of data [Barnett & Lewis1994], or which deviate so much from other observations as to arouse suspicions that they were generated by a different mechanism [Hawkins1980]. In classical data mining (i.e., mining of non-spatial data), the identification of global outliers can lead to the discovery of unexpected knowledge and has a number of practical applications in areas such as detection of credit card fraud and voting irregularities, athlete performance analysis, and severe weather prediction. These classical techniques however do not work well in the spatial domain. Spatial outlier detection techniques focus on observations of spatial data that appear to be inconsistent with their neighborhoods. Detecting spatial outliers is useful in many applications of geographic information systems and spatial databases. Such techniques can be applied for automating the map error detection step of the MSDS refinement process.

**Spatial Co-Location:** Co-location patterns represent subsets of Boolean spatial features whose

instances are often located in close geographic proximity. Examples include symbiotic species, e.g. the Nile crocodile and Egyptian plover in ecology and frontage roads and highways in metropolitan road maps. Boolean spatial features describe the presence or absence of geographic object types at different locations in a two-dimensional or three-dimensional metric space, e.g., the surface of the Earth. Examples of Boolean spatial features include plant species, animal species, road types, cancers, crime, and business types. Co-location rules are models to infer the presence of Boolean spatial features in the neighborhood of instances of other Boolean spatial features. Spatial co-location techniques discover co-location rules from a given data set. These techniques would be useful for attribute prediction as well as for error detection during the MSDS refinement process. Consider the discovery of a co-location rule that finds that all bridges co-located with populated cities are of a certain load-bearing category. This rule may be used for predicting attributes of bridges in cities when this attribute is missing in the source data set.

**Spatial Clustering** techniques group a set of spatial objects into clusters so that objects within a cluster have high similarity in comparison to one another, but are dissimilar to objects in other clusters. For example, clustering is used to determine the “hot spots” in crime analysis and disease tracking. Hotspot analysis is the process of finding unusually dense event clusters across time and space. Such hotspot analysis may be particularly useful in refining MSDS for urban warfare needs.

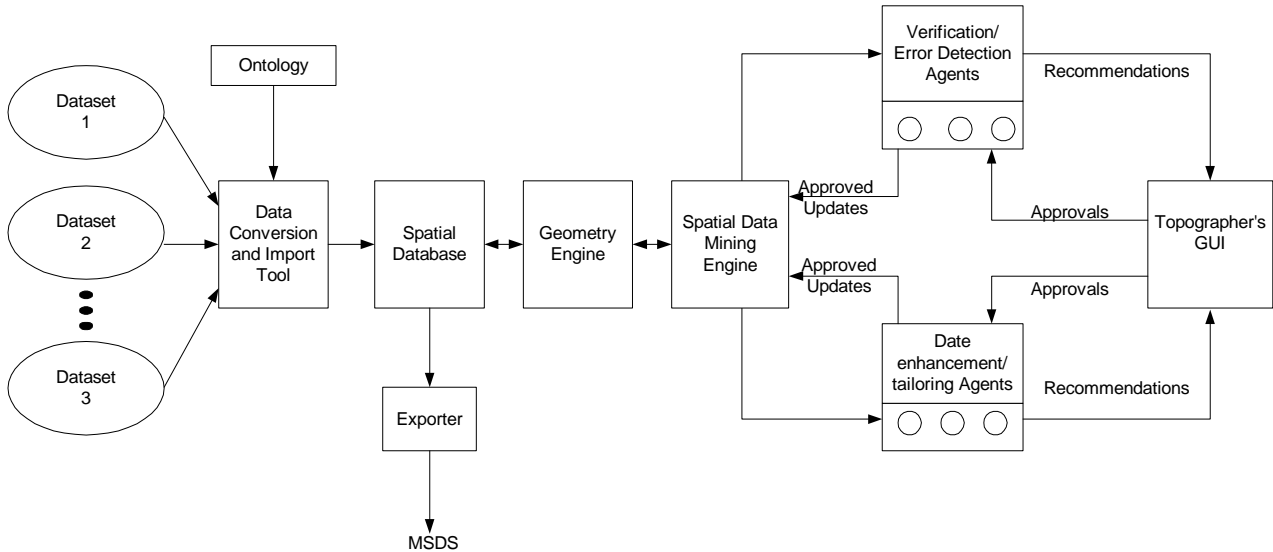
#### **Open Extensible TopoAssistant Architecture**

Automated spatial data mining techniques will be used within TopoAssistant as part of a two phased operation for discovering or predicting the needed features and attributes for a MSDS as well as for verifying/deconflicting the dataset. During the first phase, the spatial data mining techniques will process a reference or training dataset that is representative of the target MSDS. During this step, the spatial data mining techniques will automatically discover implicit patterns or rules governing the relationship among spatial objects in the training dataset. The latter may potentially be derived from diverse data sources.

The automated discovery of relationships or rules will be obtained by examining the data from different views. A set of rules may, for example, be discovered by mining co-incidental objects from different thematic layers (e.g., transportation, hydrography), i.e., a vertical view of the dataset. An example of such a vertical relationship that may be discovered is one that states that a bridge always exists where a vector road intersects a vector river. A different set of rules may be derived by mining the data for horizontal or adjacency relationships among objects. An example of such a rule is one that posits that bridges close to regions with mines or heavy industries are of a certain load bearing class. At the end of this first phase, the set of discovered rules will be presented to the user of the tool (i.e., the topographer) who is given the option of accepting or rejecting any of the rules.

During the second phase of its operation, TopoAssistant will apply the accepted set of rules to the target area for the MSDS to discover needed features and attributes and to perform verification of the features in the dataset. The discovered features, attributes, and anomalies will then be presented to the topographer with supporting rationale for each discovery. For instance, TopoAssistant may apply the first rule above to discover a set of bridges in the dataset. It may then apply the second rule to discover the load class attribute of the bridges. The topographer may then approve any subset of the feature enhancements and corrections suggested by TopoAssistant. Approved enhancements will then be applied to the target dataset to refine, enhance, or correct it.

Figure 2 shows the conceptual architecture of the proposed TopoAssistant toolkit. The figure shows the major functional components of the architecture and the information flows between



**Figure 2: TopoAssistant Conceptual Architecture**

these components. The **Data Conversion and Import** component of TopoAssistant will be driven by a user-specified ontology that will be used to convert data sets with potentially different data types into a common data model that is stored within the **Spatial Database**. The **Spatial Data Mining Engine** is the “hub” of the system. Spatial outlier detection, spatial collocation, location/attribute prediction and spatial clustering algorithms will be implemented in the spatial data mining engine. The **Geometry Engine** provides implementation of OpenGIS (OGIS) functions like distance, intersection etc [SQL3/OGIS 2004] that are leveraged by the spatial data mining engine to compute spatial joins that are necessary to find the spatial patterns. Two sets of mining agents, the **Data Enhancement/Tailoring Agents** and the **Verification/Error Detection Agents**, will operate on the source data set stored in the Spatial Database. These agents will leverage spatial data mining techniques in the spatial data mining engine to perform feature and attribute refinement, and error detection on the source data.

Different agents may focus on different aspects of the task. For instance, a verification/error detection agent leveraging the spatial outlier detection techniques may just focus on finding discrepancies like positional error and topological errors on maps. Similarly, different data enhancement/tailoring agents may focus on different subsets of the features and attributes of the source data set in performing refinement of this data set. Data enhancement/tailoring agents will leverage implementations of spatial collocations and location/attribute prediction techniques to discover missing categorical and numerical attribute values. The findings or recommendations of the mining agents will be presented to the topographer through the **Topographer’s GUI**. Recommendations that are approved by the topographer will then be used by the mining agents to update the spatial database with the newly discovered features and attributes or to correct an erroneous feature or attribute in the data set.

Once the mining agents have completed operating on the database, the **Exporter** will enable the resultant MSDS to be produced in the desired format for export to an application needing this data set. The agent based approach for implementing the data mining applications within TopoAssistant makes this tool very extensible. New applications of spatial data mining can easily be incorporated by implementing a new agent that can then be added into the system.

#### **1.4 Summary of the Phase I Results**

To establish technical feasibility we focussed on evaluating concept feasibility and implementation feasibility of the TopoAssistant approach. Concept feasibility was evaluated by using a benchmark dataset from TEC with half a dozen layers and numerous features describing a region in Korea. Spatial data mining techniques were able to identify interesting, useful and non-trivial patterns relating to MSDS verification and densification. These patterns were reviewed by TEC experts for usefulness in the MSDS refinement process by comparing the discovered patterns against auxiliary map layers.

Implementation feasibility was evaluated by designing an extensible architecture (Figure 2), implementing a concept demonstration prototype (Figure 3), and evaluating its computational performance. The computational performance (see Tables 4 & 5) of the concept demonstration prototype on the benchmark TEC dataset exceeded expectations. The performance measurements also helped reveal the computational bottlenecks that may need to be addressed to refine and verify large MSDS. In addition, this prototype was successfully demonstrated to the TEC COTR on April 26<sup>th</sup> 2004 during a visit to ATC facilities. The prototype was successfully demonstrated to FBKFF/Warriors Edge group in ARL Aberdeen on May 18<sup>th</sup> and 19<sup>th</sup> 2004.

#### **1.5 Report Outline**

Section 2 describes the results obtained during the Phase I effort. Section 3 describes the related work. Section 4 describes our commercialization approach for TopoAssistant. Section 5 presents the conclusions.

## **2. Phase I Results**

### **2.1 Phase I Prototype Implementation Approach**

The TopoAssistant Phase I prototype implementation (Figure 3) is divided into 4 components.

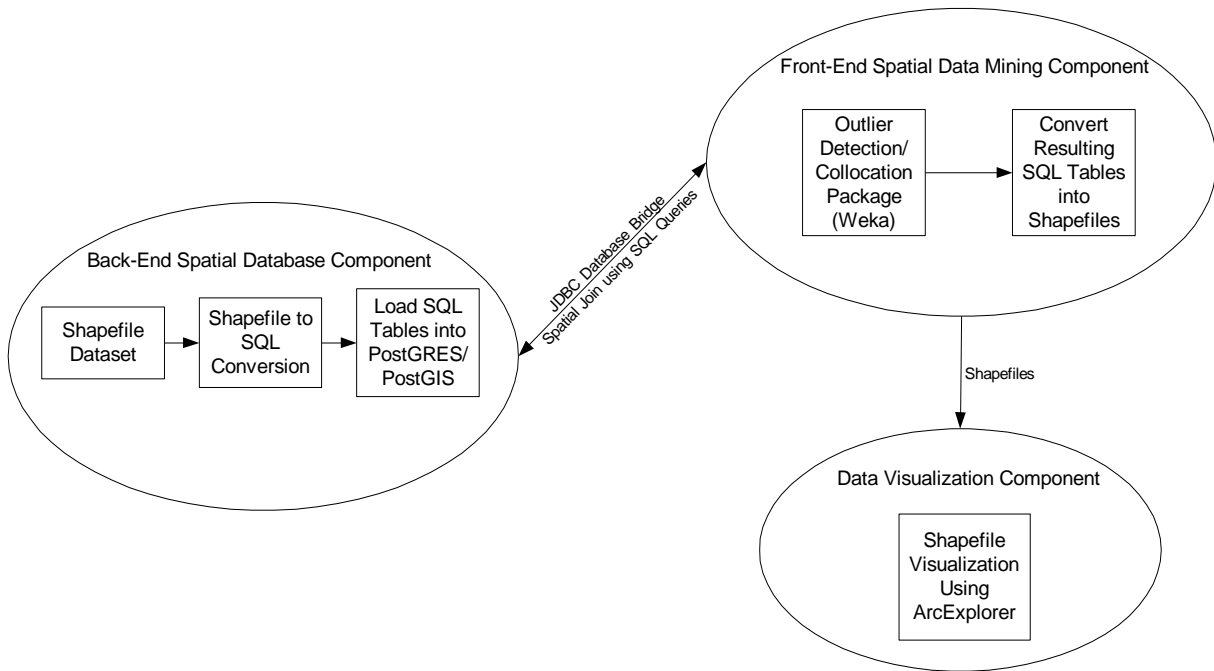
1. Back-end Spatial Database Component
2. JDBC Database Bridge
3. Front-end Data Mining Component
4. Data Visualization Component

All the components in the Phase I prototype were implemented using public domain open source software on the Linux operating system.

1. Back-end Spatial Database Component

As depicted in Figure 6, the back-end spatial database component converts ESRI shapefiles into Sequential Query Language (SQL) tables. These SQL tables are then progressively inserted into a spatially enabled PostGIS [PostGIS 2004] database in Postgres [Postgres 2004]. The spatially enabled PostGIS database contains an additional geometry column in which simple geometries such as point, line, linestring and polygons can be stored. GIS data such as shapefiles can also be easily loaded into PostGIS. The *shp2pgsql* data loader tool (available with PostGIS) converts ESRI shapefiles into SQL suitable for insertion into a PostGIS/PostgreSQL database. Once the GIS data has been converted into SQL, SQL queries can be constructed to do comparisons and

queries on the spatial data. Operations such as select and project can be constructed in SQL to extract information from the spatial tables. In addition, operations like spatial joins can also be easily performed using SQL in conformation with OGC standards (Simple feature specifications for SQL [SQL3/OGIS 2004]). Spatial join queries implemented in SQL are leveraged for map error detection and feature attribution. The join queries are executed using the Postgres and PostGIS query engines.



**Figure 3: Phase I Prototype Implementation**

## 2. JDBC Bridge

The Java Database Connectivity (JDBC) bridge enables us to construct a Java client in the open source Weka data mining tool (Witten& Eibe 1999) which can connect and query spatially enabled databases in PostGIS/Postgres. The JDBC bridge also enables us to construct spatial join SQL queries using Java code. Java clients in Weka can access PostGIS “geometry” objects in Postgres database using JDBC extensions bundled with PostGIS.

## 3. Front-end Data Mining Component

We have incorporated new Java packages into the Weka data mining software called spatial outlier and spatial collocation. These packages use spatial data mining techniques written using spatial join queries to discover positional and topological errors as well as to predict extra, erroneous or missing features in a spatial dataset. Connectivity with the back-end spatial database component is achieved via the JDBC bridge.

## 4. Data Visualization Component

SQL tables that have been returned as a result of the spatial join queries applied to discover spatial outliers are converted into shapefiles using a tool available with PostGIS called *pgsql2shp*. This helps us to visualize the vector maps with the spatial outliers and the collocation patterns.

## 2.2 Results Obtained by Applying Phase I Prototype to TEC Dataset

During Phase I, the concept feasibility of the TopoAssistant approach was established by using novel spatial data mining techniques for verification/error detection and data enhancement/tailoring of a spatial dataset provided to us by TEC. Implementation feasibility of our approach was established by developing a concept demonstration prototype that leveraged these spatial data mining techniques. In the following paragraphs, we describe the TEC data set and the results obtained by the application of our Phase I prototype to the data set.

### TEC Dataset Description

The TEC dataset is vector data for a mountainous area in Korea. The latitude for this data set ranges from 37deg 15min to 37deg 30min and the longitude ranges from 128deg 23min 51sec to 128deg 43min 52sec.

Layers in Korea Dataset	Features in Each Layer
Transportation	Roads, Cart tracks, Railways, Bridges
Surface Drainage	River/stream, Canal, Island, Common open water, Ford, Dam
Vegetation	Cropland, Rice field, Land subject to inundation, Evergreen trees, Mixed trees
Obstacles	Cut, Embankment, Depression
Soils	Poorly graded gravel, Clayey sand, Organic silt, Disturbed soil

**Table 1: TEC Korea Dataset – Important Layers and Features**

Table 1 lists the important layers and the features present in each layer in the TEC Korea data set.

Table 2 lists the results obtained by the application of spatial data mining techniques to the TEC Korea data set.

MSDS REFINEMENT TASKS	MSDS REFINEMENT TASKS (GRANULAR)	SPATIAL DATA MINING (SDM) TECHNIQUE USED	RESULTS OBTAINED VIA APPLICATION OF SDM TO TEC KOREA DATASET
Error detection & Verification	Identifying Potential Positional and Topological Errors	Spatial Outliers - Statistical/Empirical Rules	1. Disconnected Roads Detected 2. Road Frequently Crossing River Detected
	Identifying Potential Extra/Erroneous/Missing Features	Spatial Outliers – Collocation Rules	1. Cropland Outliers Detected
Data Enhancement & Tailoring	Identifying Potential Missing Features (Categorical)	Location/Attribute Prediction – Collocation Rules	1. Missing Road/River Features Detected 2. Road Collocated with River/Stream Detected

**Table 2: Results Obtained by Application of SDM to TEC Korea Dataset**

We describe the various spatial patterns discovered during the Phase I effort in this final report.

### **Error Detection and Verification**

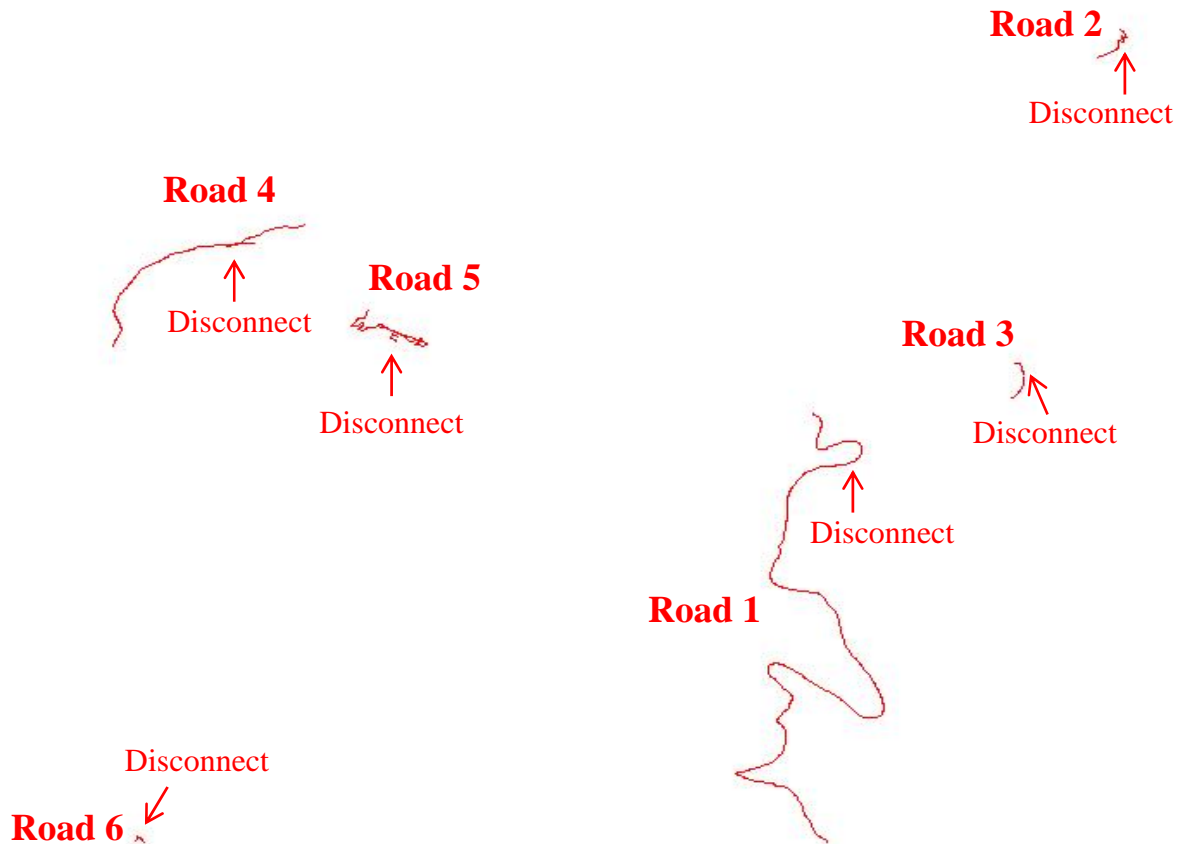
The error detection and verification phase of the MSDS refinement process involves

1. Identifying potential positional and topological errors of features on the map
2. Identifying potential extra/erroneous/missing features on the map

#### 1. Identifying potential positional and topological Errors

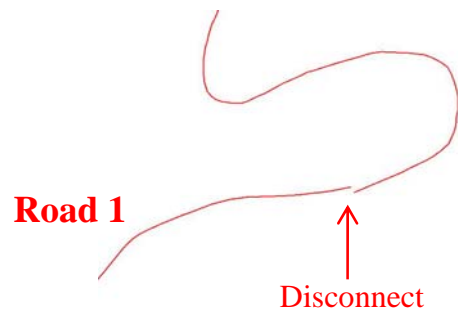
##### Disconnected Road Outlier

Figure 4 shows all the disconnected roads automatically discovered by ATC's Phase I concept demonstration prototype in the TEC Korea dataset. Disconnected road outliers may be representative of potential positional errors (roads in wrong position) or topological errors (disconnected road geometries). Disconnected road outliers were detected using statistical/empirical based spatial outlier detection techniques.



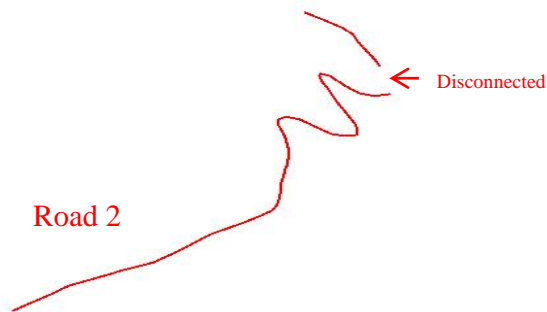
**Figure 4: Disconnected Roads Detected in TEC Korea Dataset**

This error is non-trivial because visually detecting this error through manual inspection is very time consuming. It is critical to detect and correct these errors to enable the use of automated tools that operate upon these datasets to generate TDAs or perform route planning.



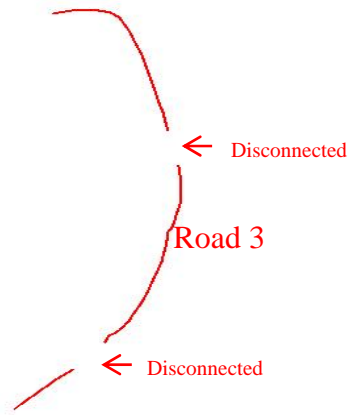
**Figure 5: Zoomed in View of Disconnected Road 1**

Figure 5 is a zoomed in view of disconnected Road 1 from Figure 4 to show the actual disconnect.



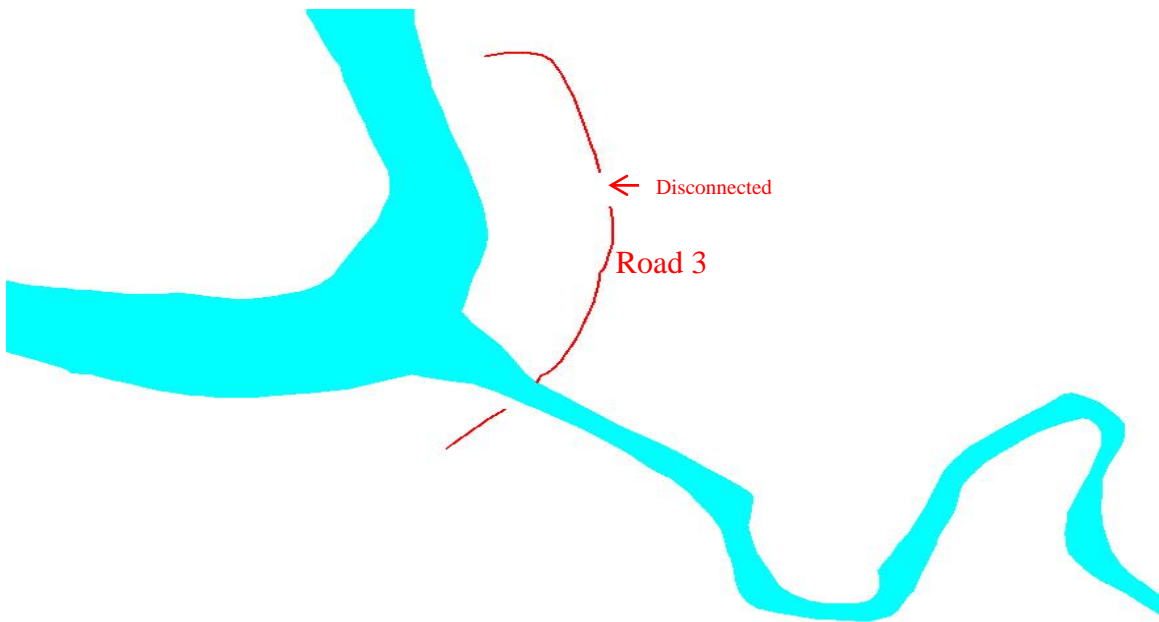
**Figure 6: Zoomed in View of Disconnected Road 2**

Figure 6 is a zoomed in view of disconnected Road 2 from Figure 4 to show the actual disconnect.



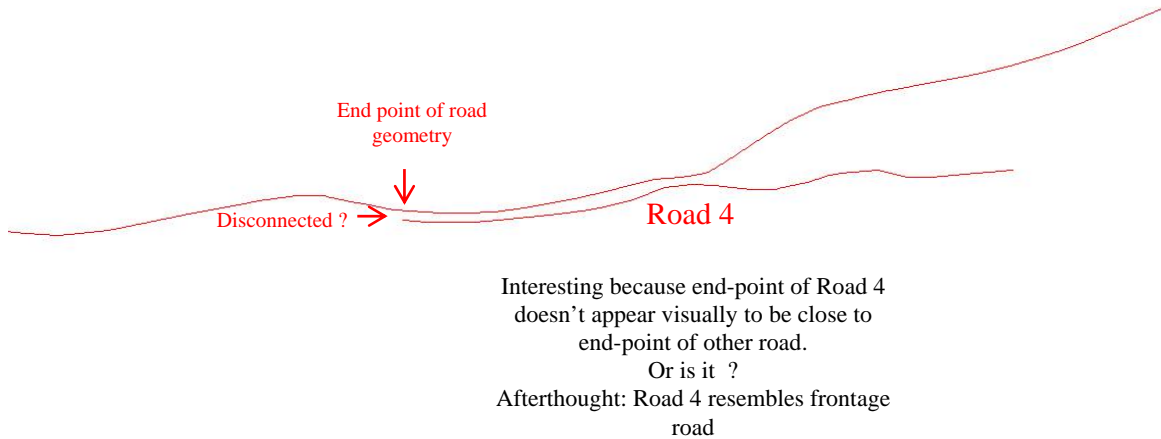
**Figure 7: Zoomed in View of Disconnected Road 3**

Figure 7 is a zoomed in view of disconnected Road 3 from Figure 4 to show the actual disconnect.



**Figure 8: Zoomed in View of Disconnected Road 3 With River Layer**

Figure 8 is a zoomed in view of disconnected Road 3 from Figure 4. The figure shows that one of the disconnects in road 3 is because of a river crossing the road.



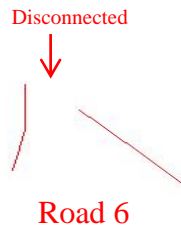
**Figure 9: Zoomed in View of Disconnected Road 4**

Figure 9 is a zoomed in view of disconnected Road 4 from Figure 4 to show the actual disconnect. On closer observation it can be noticed that road 4 is not disconnected. Road 4 resembles a frontage road.



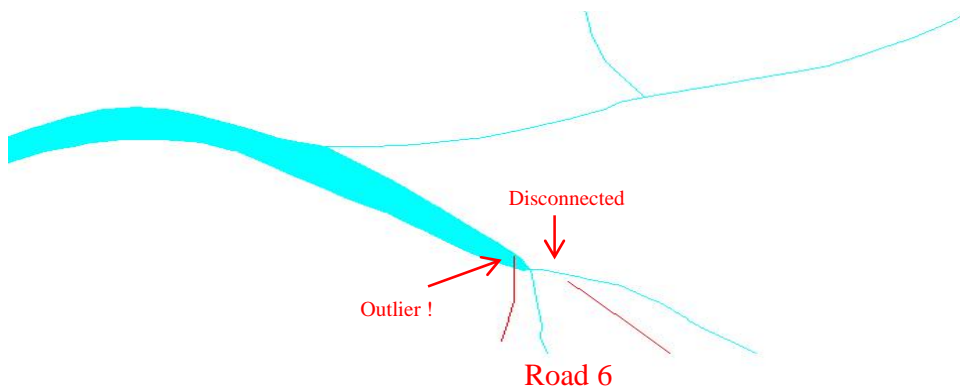
**Figure 10: Zoomed in View of Disconnected Road 5**

Figure 10 is a zoomed in view of disconnected Road 5 from Figure 4 to show the actual disconnect.



**Figure 11: Zoomed in View of Disconnected Road 6**

Figure 11 is a zoomed in view of disconnected Road 6 from Figure 4 to show the actual disconnect.



**Figure 12: Zoomed in View of Disconnected Road 6 With River Layer**

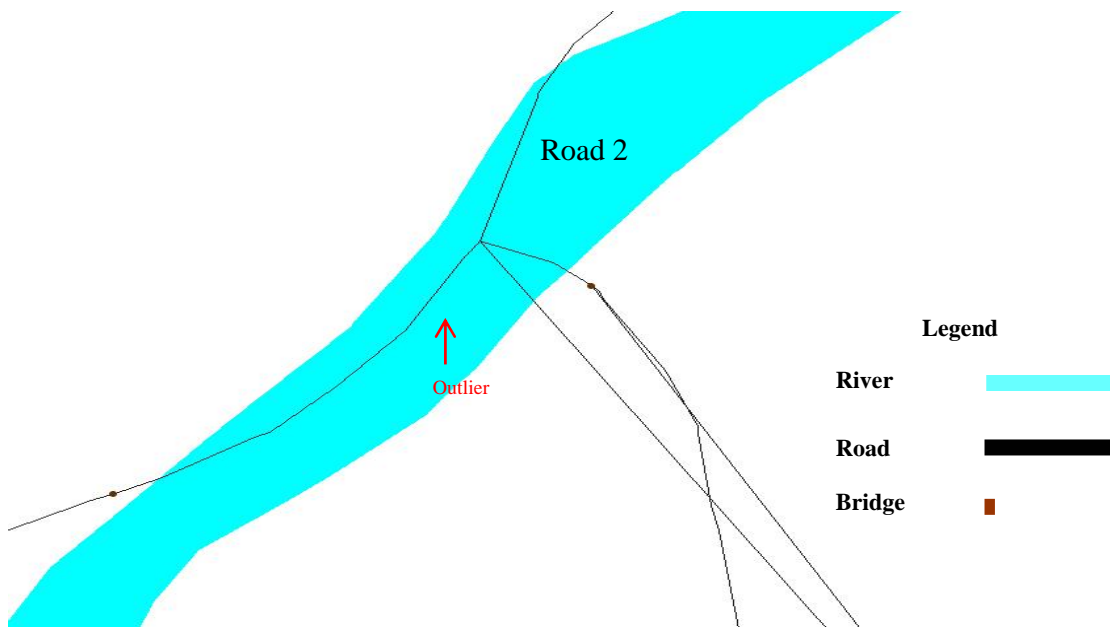
Figure 12 is a zoomed in view of disconnected Road 6 from Figure 4 with the river layer. The figure shows that road 6 leads into the river. Roads arbitrarily leading into bridges is also an outlier.

Detection of the disconnected road outlier was accomplished using spatial queries. Initially the start-point and end-point of each road is determined and stored. Roads whose start and end points are at a distance less than 0.001 units from any other road are flagged as outliers. In addition to this, we also ensure that there are no other road geometries or cart roads joining the disconnected road.

## 2. Identifying potential extra/erroneous/missing features on the map

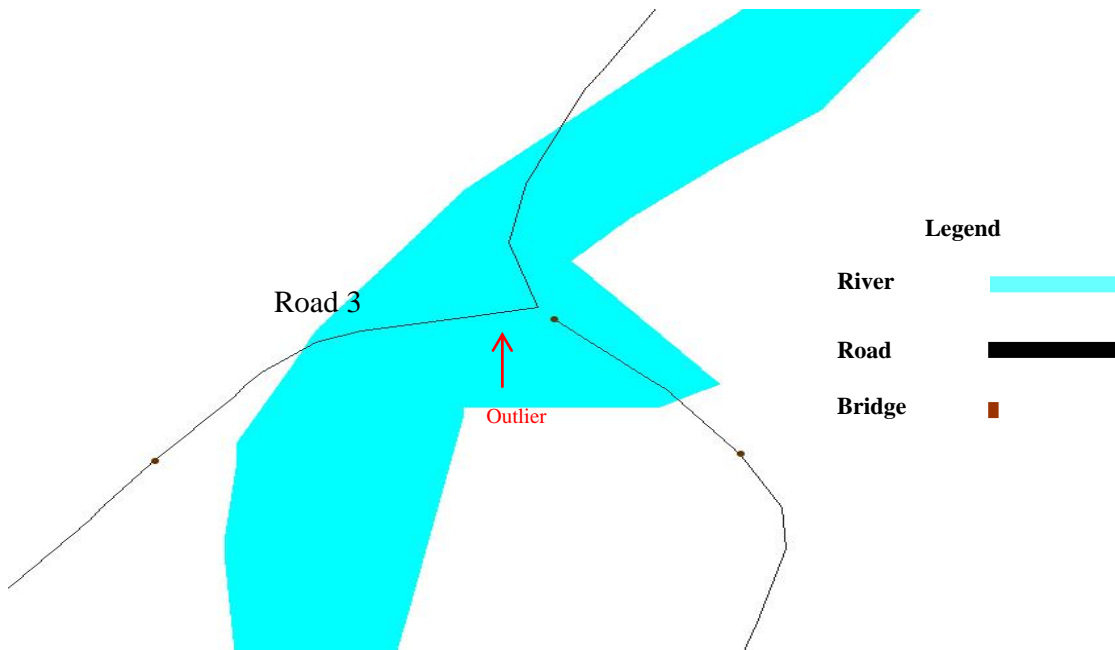
### Road Frequently Crossing River Outlier

Figure 1 depicts the road frequently crossing river pattern that was discovered automatically by our Phase I concept demonstration prototype. In the figure, roads cross rivers in some locations with no bridges in the vicinity. This spatial pattern indicates positional error and/or missing features on the map.



**Figure 13: Road Frequently Crossing River**

Figure 13 depicts another road frequently crossing river pattern that was discovered automatically by our Phase I concept demonstration prototype.



**Figure 14: Road Frequently Crossing River**

Figure 14 depicts another road frequently crossing river pattern that was discovered automatically by our Phase I concept demonstration prototype.

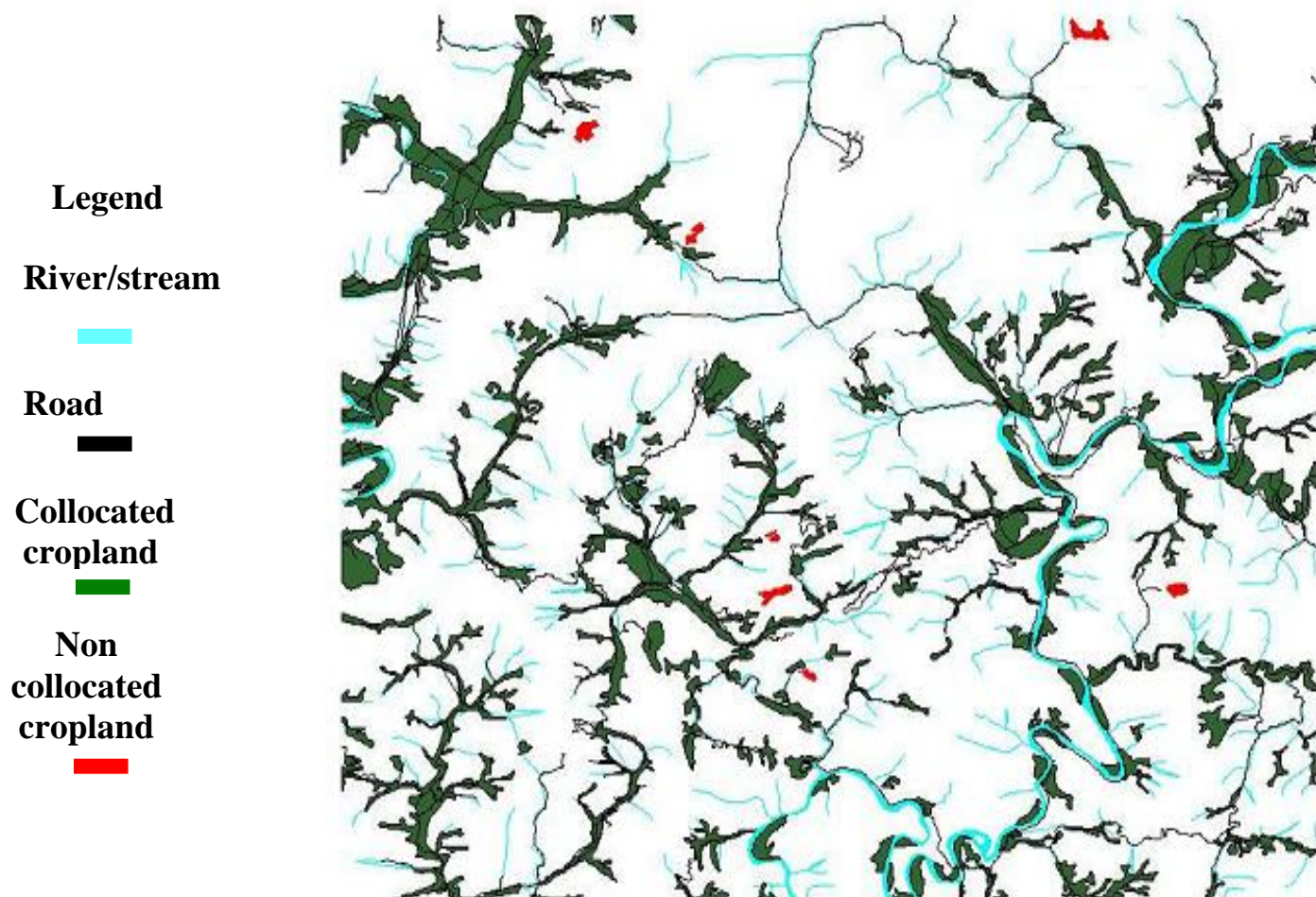
This spatial pattern was detected using statistical/empirical spatial outlier detection techniques. Initially location pairs representing intersections of roads and river are determined. If the distance between any two of these location pairs is less than 0.001 units and there is no bridge geometry between them, it is classified as an outlier.

### **Data Enhancement and Tailoring**

#### *Identifying Potential Missing Features*

#### Cropland (including rice fields) collocated with roads and rivers pattern

Figure 15 depicts the cropland/road/river collocation pattern. This figure is best viewed in color. In the figure, cropland collocated with roads or rivers are shown in green. Cropland not collocated with roads or rivers are in red.



**Figure 15: Identifying Potential Missing Road/River Features (Best Viewed in Color)**

This may identify potentially missing road or river feature, given that 96.5 % of the croplands are collocated with transportation (road/carttrack) or surface drainage (river/stream/canal) as explained later in this section. Cropland that are not close to roads or rivers are cropland outliers and may also be indicative of positional error of cropland.

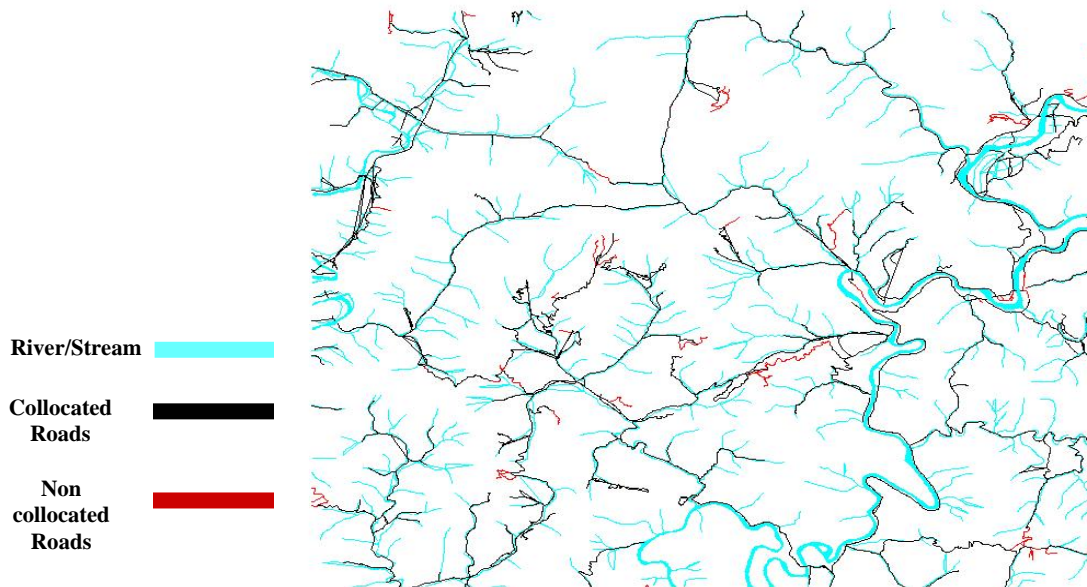
Rationale for Identifying Potentially Missing Features

To calculate the degree of collocation we compute the probability of finding a transportation or surface drainage feature nearby. Table 3 shows these probabilities for different transportation features. The first row indicates that 46% of the croplands are collocated with rivers within a distance of 0.001 units. The last row indicates that 96.5 % of the cropland locations are within a distance of 0.001 units from transportation (e.g., road, cart track) or surface drainage (e.g., river/stream/canal) features.

Collocation Pattern	Number of Collocated Cropland	Percentage of Cropland collocated with selected feature
Cropland with river	90 of 199	46 %
Cropland with cart track	97 of 199	55%
Cropland with road	118 of 199	60%
Cropland with canal/stream	137 of 199	68%
Cropland with road or cartroad or river or canal	192 of 199	96.5 %

**Table 3: Cropland/Road/River Collocation Pattern Interest Measure**

Roads Collocated with Rivers or Streams Pattern



**Figure 16: Roads Collocated with Rivers or Streams (Best Viewed in Color)**

Figure 16 depicts the road/river collocation pattern. This figure is best viewed in color. In the figure, roads collocated with rivers or streams are shown in black. Roads not collocated with rivers or streams are in red.

This may identify potentially missing river or stream feature, given that 77 % of the roads are collocated with surface drainage (river/stream/canal). Roads that are not close to rivers or streams may also be indicative of positional error of the road.

### 2.3 Phase I Prototype Scalability and Performance

Table 4 depicts the execution time of the spatial query used to discover the cropland/road/river collocation pattern. All Phase I results were obtained by executing our spatial data mining techniques on a 1.4 GHz Athlon machine with 512 MB of RAM.

Collocation Pattern	Execution Time (Minutes)
Cropland with canal	6.3
Cropland with river	2.2
Cropland with cartroad	1.8
Cropland with road	3.2
Cropland with road or cartroad or river or canal	<b>13.5</b>

**Table 4: Cropland/Road/River Collocation Pattern – Performance**

Spatial Pattern Detected	Execution Time (Minutes)
Disconnected Roads	4.5
Road Frequently Crossing River	5
Cropland Collocated with Road/River	13.5
Road River Collocation	12

**Table 5: Phase I Prototype Performance**

Table 5 summarizes the performance of our Phase I prototype. The performance is satisfactory for the TEC Korea spatial dataset but may need to be tuned for larger or higher resolution MSDS. Performance can also be considerably enhanced by using modern machines with faster CPU's and larger RAM.

### 2.4 Technical Challenges Identified during Phase I

In the following sub sections we describe the technical challenges identified during the Phase I effort.

#### 2.4.1 Flexible and Compatible TopoAssistant Implementation

As described previously, the Phase I concept demonstration prototype was implemented using public domain open source software. The advantage of using public domain open source software is low-cost and easy access to the source code. The performance of spatial data mining techniques was satisfactory when implemented using open source software. But it may be possible to achieve better computational performance by leveraging proprietary COTS GIS products. In addition, ESRI's family of GIS products are also widely used by the Army and other branches of DoD. Topographers may find the TopoAssistant tool more useful if it is compatible with the commercial GIS applications like ArcGIS and ArcSDE. Therefore, the TopoAssistant software tool will be most useful to topographers if it is compatible with both open source software and commercial software.

#### 2.4.2 Scalability Challenge

1. Multi-way spatial join

Spatial joins written using SQL are used in our Phase I prototype to discover spatial patterns like collocation patterns etc. Spatial join over more than two tables can be computationally expensive. For example, to find the cropland areas collocated with roads and rivers, a spatial join can be performed over five tables simultaneously. The five tables are the road table, cart track table, river table, stream/canal table and the cropland/rice field table. Spatial joins over more than two large tables usually take a few hours to execute. Techniques need to be designed that can speed up the spatial join process.

## 2. Top-k querying technique

Many spatial data mining techniques identify a large number of patterns even though a topographer may be interested in a small number, (say k) patterns due to the constraints of time and other resources. A key challenge for current spatial data mining techniques is to exploit this information to scale up to larger MSDS by eliminating the effort to identify unnecessary patterns beyond the top k patterns requested by the topographer.

### 2.4.3 Modeling Spatial Patterns

#### 1. Local Spatial Outliers

Statistical tests for detecting global outliers may not identify spatial outliers. Spatial outliers are significantly different from their neighborhood even though they may not be significantly different from the entire population. For example, a brand new house in an old neighborhood of a growing metropolitan area is a spatial outlier. Figure 17 shows a spatial outlier in the bottom left corner. A river segment highlighted via the enclosing red box has stream segments upstream and downstream. In other words, the highlighted river segment is quite different from its neighbors, even though it is not different from the overall population of streams and rivers. This may indicate potential mislabeling of this river segment or some of its neighboring stream segments, unless it is a result of a river flowing through a narrow and deep valley.

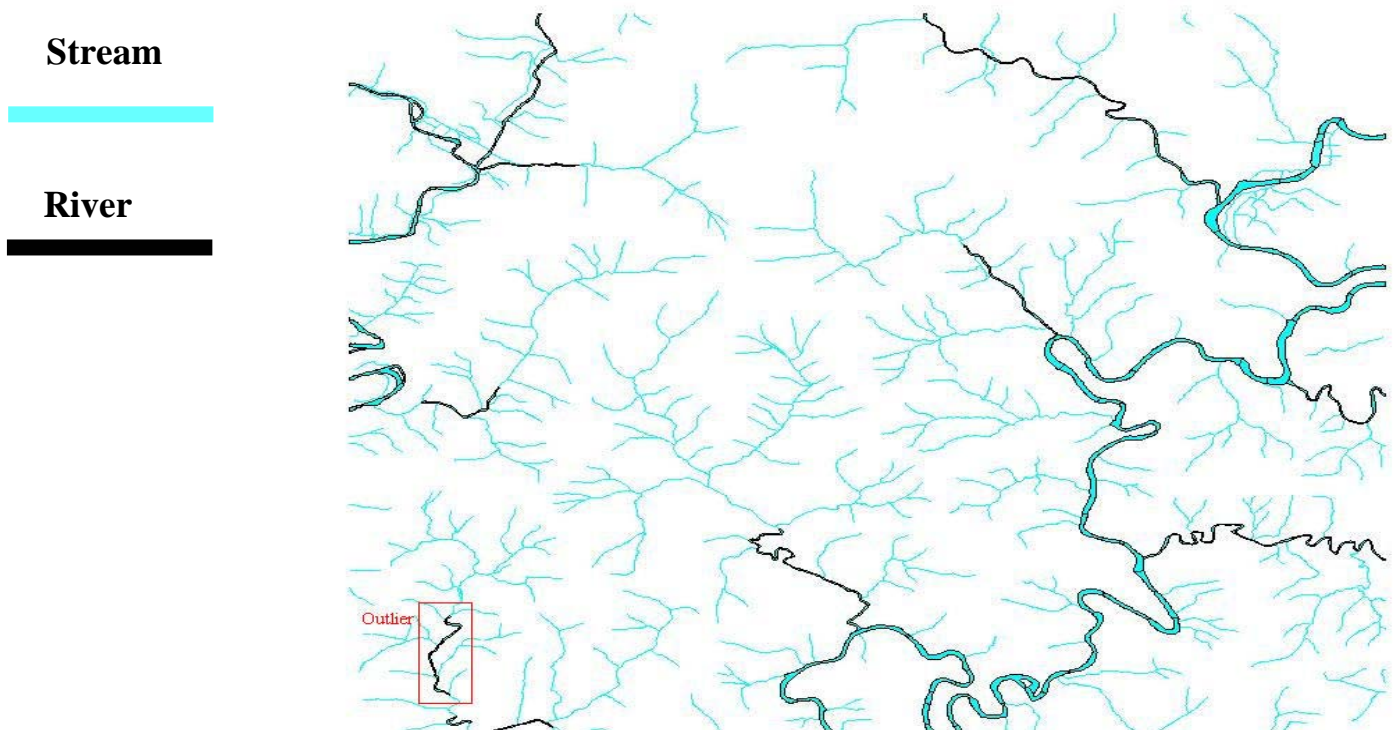


Figure 17: Local Spatial Outliers

### 3. Spatial Edge Effect

It is well known in spatial statistics that effectiveness of learned rules and models from a spatial dataset diminishes in areas close to the edge of a geographic region under study. This is broadly referred to as the spatial edge effect. Figure 18 (best-viewed in color) shows seven out of about 199 croplands in red to indicate that those are not collocated with transportation or surface drainage features. Notice that one of these is located close to the top edge near the top right corner. It is possible that a road or a river may be close to this cropland but is outside this map. Spatial data mining techniques need to account for the edge effect.

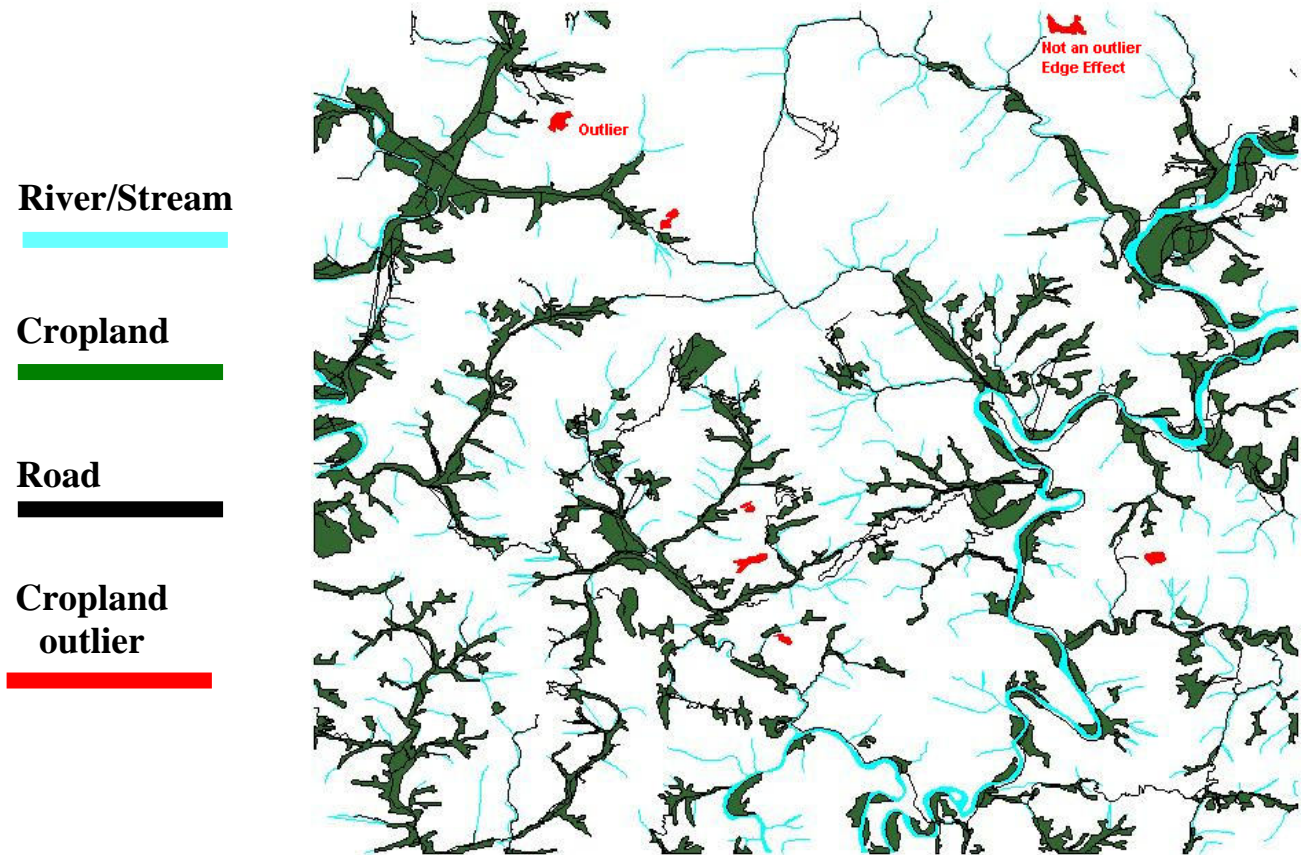
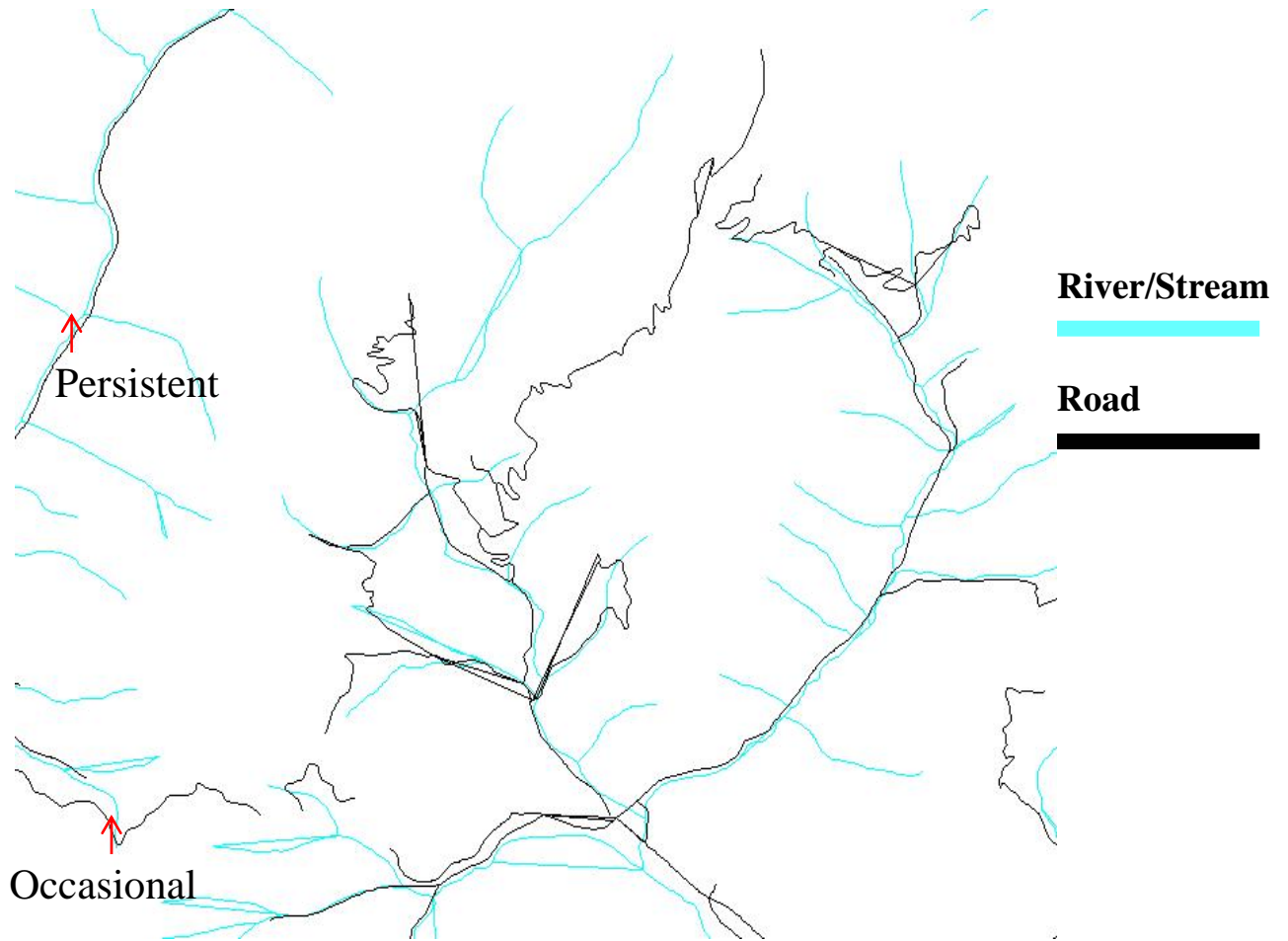


Figure 18: Spatial Edge Effect (Best Viewed in Color)

### 4. Similarity measure for extended objects

Similarity measures for point objects, e.g. distance, may not be effective for measuring similarity between extended spatial objects such as linear or polygonal features. Figure 19 shows roads that are collocated with rivers within a distance of 0.001 units. Notice that some roads are persistently collocated with rivers. An example is highlighted in the top left corner. Other roads are occasionally collocated with rivers as illustrated in the bottom left corner. Spatial data mining techniques need new interest measures to distinguish between persistent and occasional collocation of extended features.



**Figure 19: Similarity Measures for Extended Objects – Occasional versus Persistent Collocations (Best Viewed in Color)**

## 5. Spatial Heterogeneity

Many geographic patterns and models are regional rather than global. For Example TopoAssistant discovered that almost all roads in Korea dataset were collocated with rivers or streams. This pattern may be prevalent in hilly or mountainous terrain. However, this pattern may not be prevalent in dense urban areas on flat terrain. So it is important for spatial data mining techniques to identify the scope of generalizability of the discovered patterns from given regional or global spatial datasets.

### 2.4.4 Automation of spatial data mining methods

#### 1. Reduce user burden to specify thresholds

Table 6 shows that 77% of the roads in the TEC Korea dataset were found to be collocated with rivers or streams when we used a distance threshold of 0.001 units (If road is at a distance  $< 0.001$  units from a river, then the Phase I prototype reports that the road is collocated with the river).

Collocation Pattern	No of Collocated Features	Interest Measure (Collocated roads/Total roads) * 100
Road with canal/stream	153 of 239	64 %
Road with river	96 of 239	40 %
Road with stream or river	176 of 239	74 %
Cartroad with stream	97 of 136	71 %
Cartroad with river	44 of 136	32 %
Cartroad with stream or river	111 of 136	82 %
All roads with river or stream	287 of 375	<b>77 %</b>

**Table 6: Interest Measure for Road River Collocation Patterns with Distance Threshold of 0.001 Units**

Table 7 shows that 100 % of roads were collocated with rivers or streams when using a distance threshold of 0.01 units.

Collocation Pattern	No of Collocated Features	Interest Measure (Collocated roads/Total roads) * 100
Road with canal/stream	133 of 239	56 %
Road with river	238 of 239	99 %
Road with stream or river	239 of 239	100 %
Cartroad with canal/stream	130 of 136	96 %
Cartroad with river	79 of 136	58 %
Cartroad with stream or river	136 of 136	82 %
All roads with river or stream	375 of 375	<b>100 %</b>

**Table 7: Interest Measure for Road River Collocation Patterns with Distance Threshold of 0.01 Units**

Therefore, it is very important that the topographers choose the right threshold when using spatial data mining techniques to discover patterns. In fact, the burden on the domain expert topographer will be minimized if parameter free methods are designed to eliminate the need for various thresholds during discovery of spatial patterns.

## 2. Reduce user burden to enumerate patterns

The space of candidate patterns can be extremely large based on the number of map layers, the number of features within each layer, the number of spatial data mining algorithms and the heterogeneity in the dataset. Enumeration of candidate patterns by hand can be quite tedious and may limit the use of TopoAssistant by many topographers. It is desirable to provide a candidate pattern enumeration facility to reduce topographer's burden. For example, to find the cropland/road/river collocation pattern, our Phase I prototype tried all possibilities of triplets and pairs out of cropland and all other features in the TEC Korea dataset. This process was very time consuming because individual agents had to be written specifically for each triplet and pair. Automated enumeration of candidate patterns will go a long way in speeding up the spatial data mining process.

### 2.4.5 MSDS Verification/Error Detection and Data Enhancement Agent Synthesis

The methodology that we followed to build the agents for the Phase I rapid prototype is manual and time consuming. New techniques have to be designed to facilitate quick and rapid construction of new data mining agents.

### 2.4.6 Merging Spatial Datasets from Disparate Sources

The issues in merging spatial datasets are non-trivial as datasets about a common geographic region may use different spatial frameworks. For example, the Common Open Water feature in the TEC Korea dataset was not found in the Feature and Attribute Coding Catalog (FACC). Therefore tools to understand and merge the features and attributes from disparate FACCs have to be designed before spatial data mining techniques can be applied.

### 2.4.7 Provide Actionable Information to the Topographer

The results of applying the spatial data mining techniques to the training dataset in the form of discovered rule must be provided to the topographer along with the rationale used for the discovered rule etc. Similarly errors discovered as well as feature attributions made in the target dataset by application of the discovered rule should also be presented to the topographer, along with the rationale for the error and additional information such as error context.

## 2.5 Technical Approach to Address the Technical Challenges

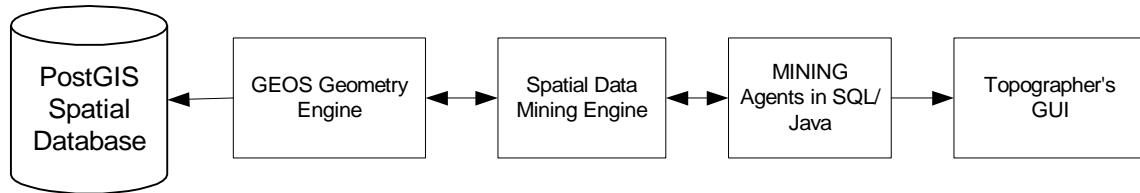
Technical Challenge	Technical Approach
1. Flexible and compatible implementation	Implementation using open source software
	Implementation using COTS and C/JMTK framework
2. Scalability Challenge	SDM Algorithms – Apriori, MRF, SAR
	Geometry Engine Level - MOBR, Convex hull, Spline
	Database Level – Spatial Indexing, Spatial Star Join
	System Level - Reduce JDBC overhead, use tables instead of views
3. Automation of Spatial Data Mining Methods	Parameter free methods - Sorting to help identify top k outliers, statistical analysis to derive thresholds
	Candidate pattern enumeration - Formulate enumeration agents to enumerate pairs, triplets, subsets of given set of features and report interesting candidates
4. Modeling Spatial Patterns	Buffer based measure for similarity measures
	Distance of pattern to edge to resolve spatial edge effect
5. MSDS Verification and Densification Agent Synthesis	SDM Process - Specify SDM process including tasks and their dependencies
	Agent composition language - Graphical/Interactive, Scripting language based
6. Merging Spatial Datasets	Ontology based solution - Universal ontology (GML),
7. Provide actionable information to Topographer	Two phase process - Rule recommendations followed by errors/feature attribution recommendation

**Table 8: Technical Challenges and Technical Approach**

Table 8 presents our technical approach for addressing the technical challenges.

### 2.5.1 Flexible TopoAssistant Implementation

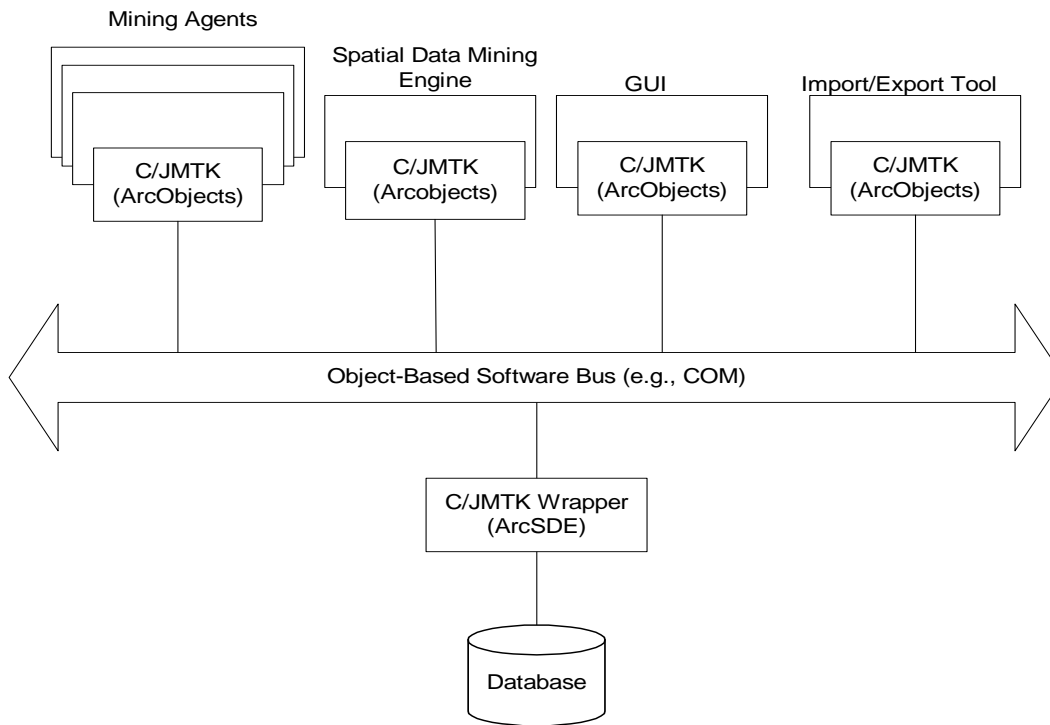
In Phase I of this SBIR, we implemented a rapid prototype to demonstrate TopoAssistant capabilities using public domain open source software. Our initial implementation of the TopoAssistant will leverage open source software components, such as the PostGIS spatial database and JDBC database bridge that were described in detail in Figure 3. Figure 20 depicts the implementation architecture for the TopoAssistant tool using open source software. We use the PostGIS spatial database to store the spatial datasets. Geometry Engine Open Source (GEOS) provides implementations of all the OGIS spatial functions such as distance, intersection, touch etc. The spatial data mining engine will contain implementations of algorithms for spatial outlier detection, spatial collocation, location/attribute prediction and clustering. The spatial data mining engine will be implemented in either Java or C++. The spatial data mining agents which provide the verification/error detection and intensification/densification functionality will be implemented either as standalone agents in Java or as a Java package in an open source data mining toolkit such as Weka. All communication between the spatial data mining engine, geometry engine and the PostGIS database will be accomplished using a JDBC/ODBC bridge.



**Figure 20: TopoAssistant Implementation using Open Source Components**

The topographer's GUI will leverage public domain open source software such as GRASS or QGIS which can be customized with the addition of new modules.

The open source software used for our implementation may not provide the most efficient platform for executing computationally intensive spatial join queries. Topographers may be more comfortable using the TopoAssistant tool on ESRI's Arc family of GIS products. Therefore we will also implement the TopoAssistant software using the Commercial/Joint Mapping Toolkit (C/JMTK) (Figure 21) as the integration framework for tying together all the functional components.



**Figure 21: C/JMTK Based Open Extensible Architecture**

C/JMTK builds upon a number of commercial products to provide an open interoperability framework for building geospatial application. Notable products integrated within C/JMTK include ArcSDE for the spatial database, ArcObjects for client-side interface wrappers, and GUI building components such as ArcView. Building upon the C/JMTK framework would enable TopoAssistant to leverage a number of software components, especially in the area of visualization, and thereby enable rapid development of the product. It would also facilitate seamless interoperability between the TopoAssistant and the C/JMTK based consumer applications of the MSDS products produced by the tool.

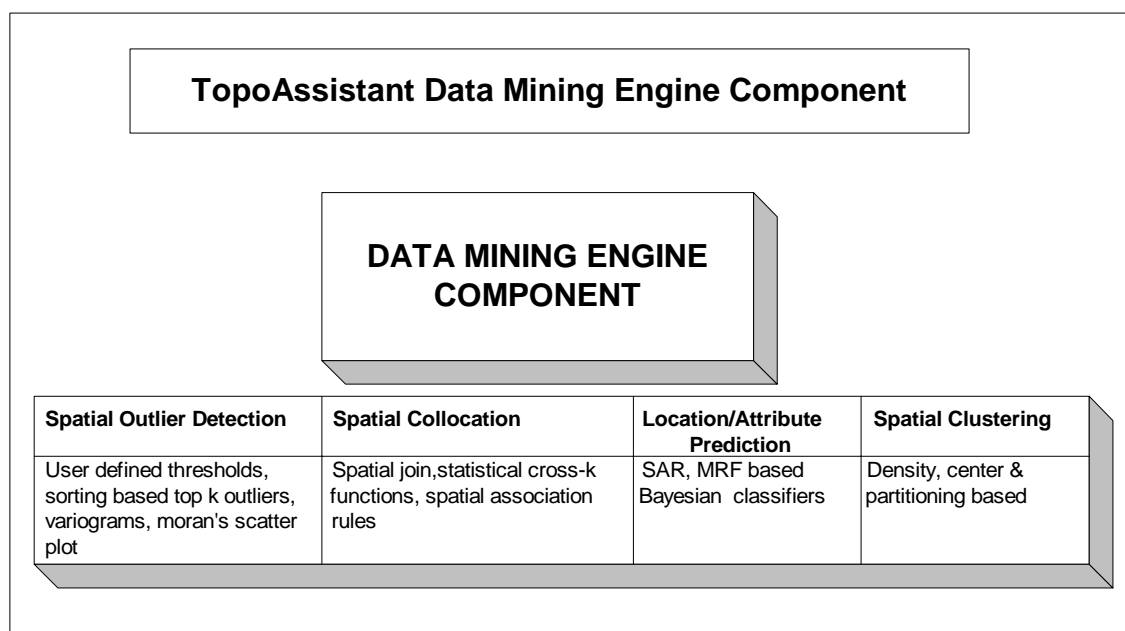
## 2.5.2 Designing SDM Level Algorithms, Geometry Engine, Database Level and System Level Optimizations

### *Spatial Data Mining Engine Design and Implementation*

Data Mining Engine Component (Figure 22) will include the following families of algorithms

1. Spatial outlier detection
2. Spatial collocations,
3. Location/Attribute prediction techniques
4. Spatial clustering

We will choose appropriate algorithms within each family and performance tune them so that they can scale up to large/fine resolution MSDS.



**Figure 22: TopoAssistant Data Mining Engine Component**

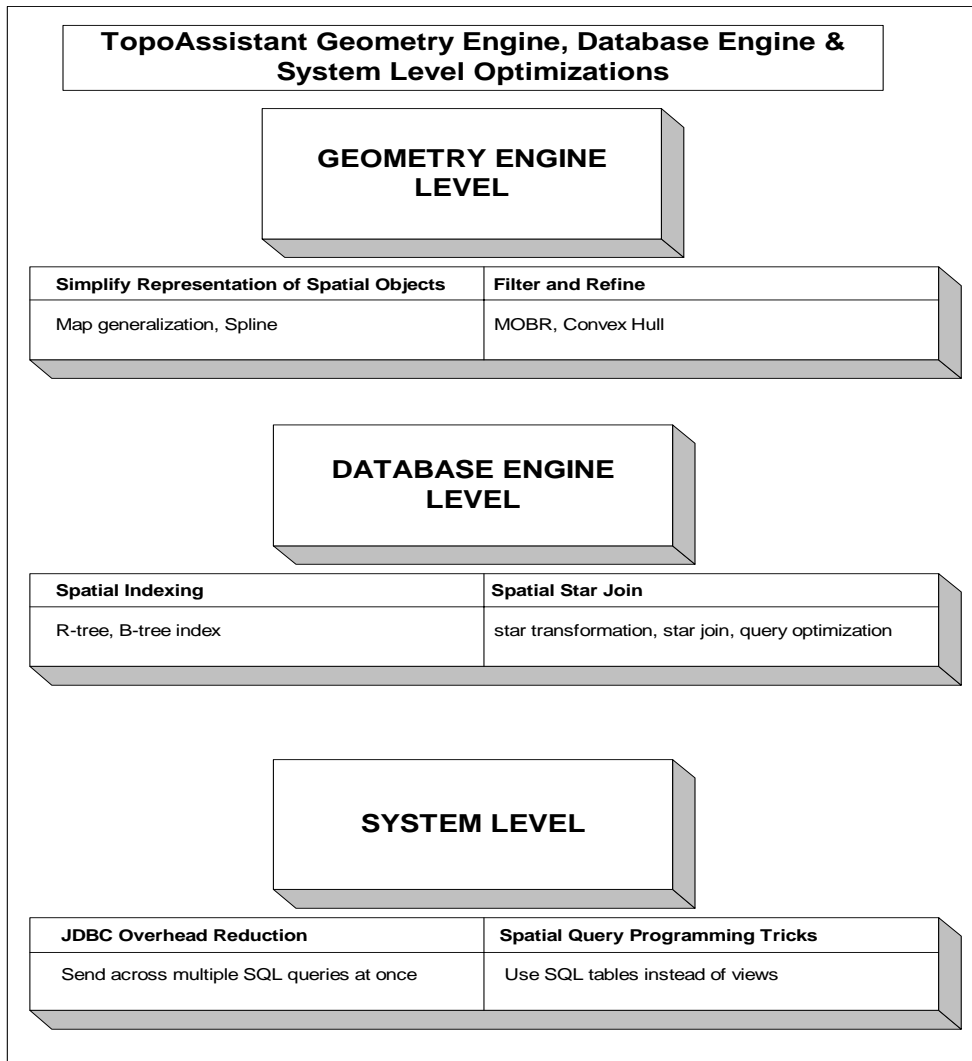
Candidate techniques for spatial outlier detection include user-defined threshold based techniques, sorting based top-k outliers, statistical techniques such as variograms and Moran's scatter plot. The computational overhead of these techniques can be reduced by using specialized data structures and algorithms. For example, many statistical techniques for spatial outliers can be implemented efficiently using spatial joins [Shekhar & Lu 2003].

Candidate techniques for collocation analysis include statistical techniques (cross-k functions), spatial association rules, and spatial join based approaches [Shekhar 2001]. The last two techniques can be speeded up by exploiting apriori principles, particularly when the interest measures are monotonic.

Candidate techniques for spatial clustering include density based, center based (k-mean, EM) and partitioning based approaches [Shekhar & Chawla2002]. Candidate techniques for location prediction include spatial autoregression (SAR) and Markov random field (MRF) based Bayesian classifiers [Shekhar et al., 2002]. The performance of location/attribute prediction techniques can be improved by using approximation methods.

**Performance Optimizations at Geometry Engine, Database Engine and System Level**

The computational performance of spatial data mining techniques can be improved by fine-tuning the parameters of the geometry engine, database engine and other system components (Figure 23).



**Figure 23: Geometry Engine, Database Engine and System Level Optimizations**

The geometry engine is used to compute similarity measures by invoking geometric operations such as distance, intersection, buffer, length, and area. The cost of these geometric operations can be reduced by **simplifying the representation** of spatial objects. For example, the geometry of a stream may be represented in the original map using a linestring with thousands of points. This linestring may be approximated by another linestring with dozens or hundreds of points for a given error tolerance using techniques such as map generalization and splines. Such an approximation may drastically reduce the cost of geometric operations while introducing a limited amount of error. A **filter and refine approach** can be used to eliminate the need of computing many expensive geometric operations such as buffer with exact geometry. For example, Minimum Orthogonal Bounding Rectangles (MOBR) or convex hulls may be used in the filter step to eliminate many objects from consideration. Exact geometries may be used in the refinement step for the remaining objects.

The Database Engine is used to compute spatial joins over multiple tables storing map features and layers. Careful selection of indices, such as R-trees or B-trees, can reduce the cost of spatial

joins [Shekhar & Chawla2002]. In addition, providing hints such as star transformation and star join can help the database query optimizers select appropriate join algorithms for a given spatial join query.

System level decisions can also impact the performance of the spatial data mining engine. For example, the collocation miner may send a large number of spatial join queries via a DBMS API such as JDBC or ODBC. The overhead of communication between the SDM engine and the DBMS can be reduced by batching multiple queries in a single communication. In addition, rewriting SQL queries can reduce their processing time. For example, nested Select statements in SQL may be rewritten as a single block select statement to improve performance.

Finally proper capacity planning in selection of computer hardware can go a long way toward improving the computational performance. The server hardware supporting the SDM engine, geometry engine, and database engine should provide adequate memory, CPU speed , I/O bandwidth, and a graphics card to support frequent spatial computations relevant to verification/error detection and intensification/densification of MSDS.

### 2.5.3 Designing Verification/Error Detection and Data Enhancement/Tailoring Agents

Two sets of mining agents - the **Data Enhancement and Tailoring Agents** and the **Verification/Error Detection Agents** - operate on the source data set. These agents implement different spatial data mining techniques to perform feature and attribute refinement, and error detection on the source data.

Different agents may focus on different aspects of the task. For instance, a Verification/Error Detection agent built using one of the spatial outlier detection techniques may just focus on finding discrepancies of a certain kind. Another Verification/Error Detection agent using another spatial outlier technique may look for another set of discrepancies. Similarly, different Data Enhancing/Tailoring agents may focus on different subsets of the features and attributes of the source data set in performing refinement of this data set. To build these agents, we will first model the spatial patterns, design techniques for automation of the spatial data mining algorithms and design a process to synthesize spatial data mining agents automatically.

#### 1. Modeling Spatial Patterns

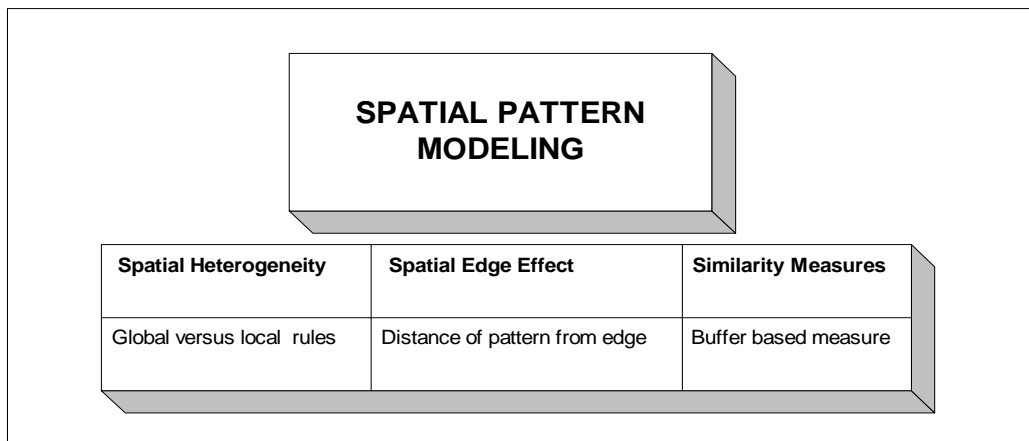


Figure 24: Modeling Spatial Patterns

Spatial heterogeneity, spatial edge effect and similarity measures for extended objects are three important issues that are addressed in the spatial pattern modeling phase (Figure 24).

Spatial Heterogeneity

It is well known from the second law of geography that many geographic patterns and models are regional rather than global. For example, rules applicable to hilly areas may not be applicable to plains. The rule of roads collocated with rivers was discovered in the Korea dataset, and an expert topographer may specify that such rules are applicable only in hilly areas). If global datasets are available for training our system may show the map of local strength of a selected collocation rule R to assist experts in identified the regional nature of R.

Spatial Edge Effect

It is well known in spatial statistics that effectiveness of learned rules and models diminishes in areas close to the edge of a geographic region under study.

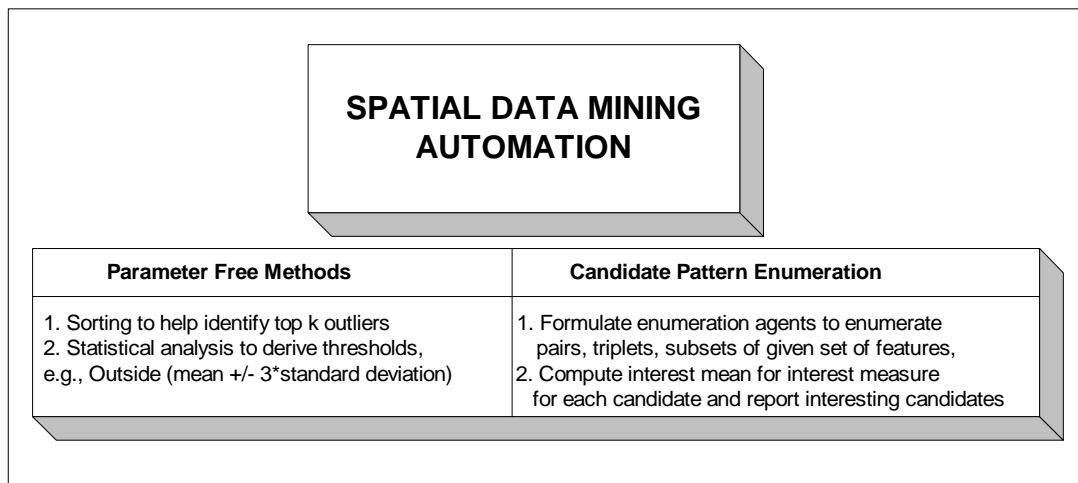
There are two ways to address this issue

- Spatial statistics model (attaching a confidence measure to each prediction, with lower confidence to predictions near the edge)
- Visual feedback - pattern distance from map edge

Similarity Measures for Extended Objects

An example measure of the strength of collocation rule R (Road implies River) is a function of average\_length (intersection(river, buffer(road, buffersize)) instead of average\_distance(river, road). The averages are computed over the collection of all roads to measure the strength of the collocation rule R. Similar measures may be needed for polygonal features using average area, intersection and buffer operations.

2. Automation of Spatial Data Mining Methods



**Figure 25: Automation of Spatial Data Mining Methods**

To automate spatial data mining techniques (Figure 25), parameter free methods and candidate pattern enumeration techniques will be designed.

### Parameter Free Methods

Many data mining techniques require users to specify values for internal parameters. Detection of disconnected road via outlier detection technique (refer section/figure) requires specification of distance thresholds. Expert users may be comfortable specifying values of such parameters. However, it may be inconvenient for many topographers for whom we would like to provide parameter free methods. This can be done via many different techniques including sorting and statistical analysis.

- Sorting based methods reorder the discovered patterns based on a specific interest measure before presenting the results to the end user. To detect disconnected roads the TopoAssistant may sort pairs of road-end-points in ascending order by distance. The topographers may browse the first several pairs based on the available time.
- Statistical methods analyze the frequency distribution of a dataset over the domain of values of an interest measure. This information may be used to identify the data points that are quite different from the population based on a default confidence threshold as well as a default choice of distribution tails. To detect disconnected roads TopoAssistant may analyze the frequency distribution of pairs of road-end-points over the domain of value of distance between road-end-points. Assuming a default confidence threshold of 99%, TopoAssistant may identify the outlying top 1% of road-end-point pairs, which are in the left tail.

### Candidate Pattern Enumeration

The space of candidate patterns can be extremely large based on the number of map layers, the number of features within each layer, the number of spatial data mining algorithms and the heterogeneity in the dataset. Enumeration of candidate patterns by hand can be quite tedious and may limit the use of TopoAssistant by many topographers. It is desirable to provide a candidate pattern enumeration facility to reduce the topographer's burden. For small MSDS, this facility may enumerate all possible combinations of map layers, map features and verification/intensification agents. However, exhaustive enumeration may be computationally exorbitant for large MSDS and TopoAssistant may employ a smart enumeration technique – such as apriori, dynamic programming, heuristic based A\* search algorithm, or genetic algorithms - to explore the most promising parts of the search space. For example, the collocation mining algorithm may enumerate various subsets of features and may use apriori principles to focus on promising subsets.

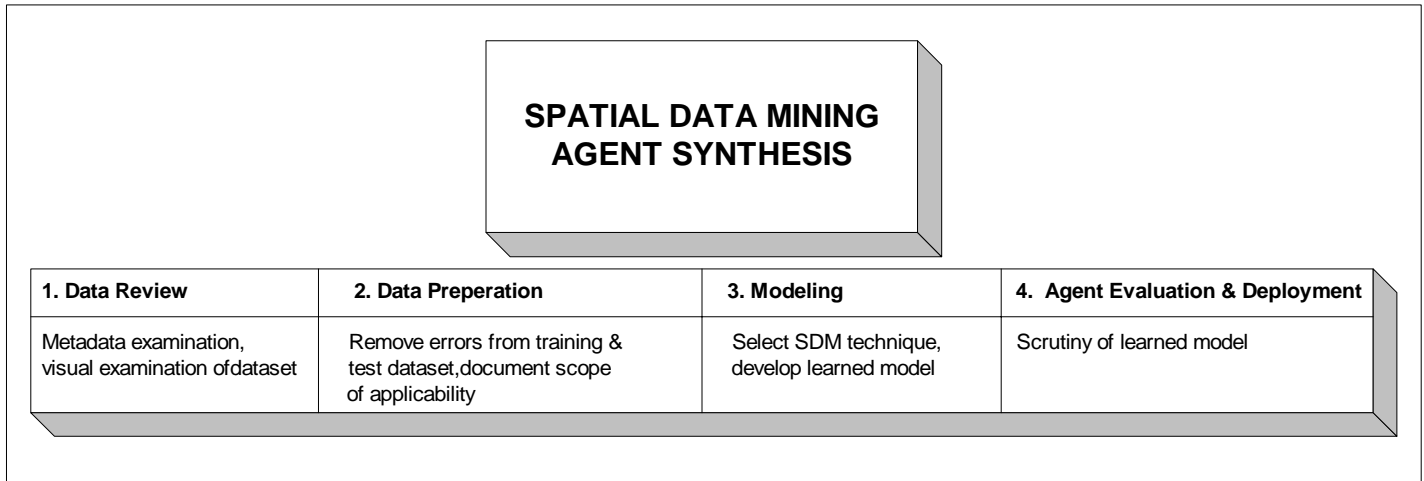
## 3. Spatial Data Mining Agent Synthesis

The main tasks in building a spatial data mining agent (Figure 26) include data review, data preparation, modeling, agent evaluation and deployment.

The data review task focuses on examination of a given MSDS using visual, querying and statistical techniques. A topographer may first examine metadata such as names and definitions of features, attribute codes and values, feature geometry data type, geographic area and timeframe, source, resolution, scale, and accuracy of attribute values. Next, the dataset may be examined by a GIS tool such ArcExplorer or GRASS.

The data preparation task focuses on selection of training and testing datasets for building and evaluating a data mining model. Outliers may be removed from training and testing datasets that will be used to build attribute value prediction models. The training and testing datasets should not only be representative of but also cover the overall datasets which will be verified/intensified

using the agent being constructed. An expert topographer may document the scope of applicability, e.g., hilly tropical areas, or deserts for the selected training and testing datasets.



**Figure 26 : Spatial Data Mining Agent Synthesis Techniques**

The modeling task is concerned with selection of a data mining technique for the given task and dataset. For example, Spatial Auto Regression (SAR) may be selected to predict a numeric attribute, while a Markov Random Field (MRF) Bayesian classifier may be preferred in predicting a categorical attribute. The selected data mining technique may be used to learn a model with the help of the learning dataset.

The evaluation task is concerned with the scrutiny of the learned model for its usefulness in analyzing new MSDS. A possible technique is to use the learned model to analyze the testing dataset and compare the actual results with the expected results. This comparison can provide an assessment of the quality of the learned model. If the model is assessed to be inadequate, an expert topographer may go back to the modeling task to explore alternative models.

If the learned model is assessed to be adequate and useful, then deployment is performed to prepare the model for other topographers. Ease of use is a major concern. Robustness in the face of unexpected situations is another concern.

During the commercialization phase we will develop tools to simplify agent synthesis by expert topographers. For example, a graphical, interactive method may be provided to synthesize an agent by extending a GUI similar to Clementine, with spatial data mining primitives. It essentially allows users to construct a small data flow graph to specify data mining tasks and their dependencies. Another option is to develop a scripting language by extending Perl, Python or ArcAML to invoke and compose SDM facilities from TopoAssistant instead of writing SQL or Java code directly.

#### 4. Spatial Dataset Merging

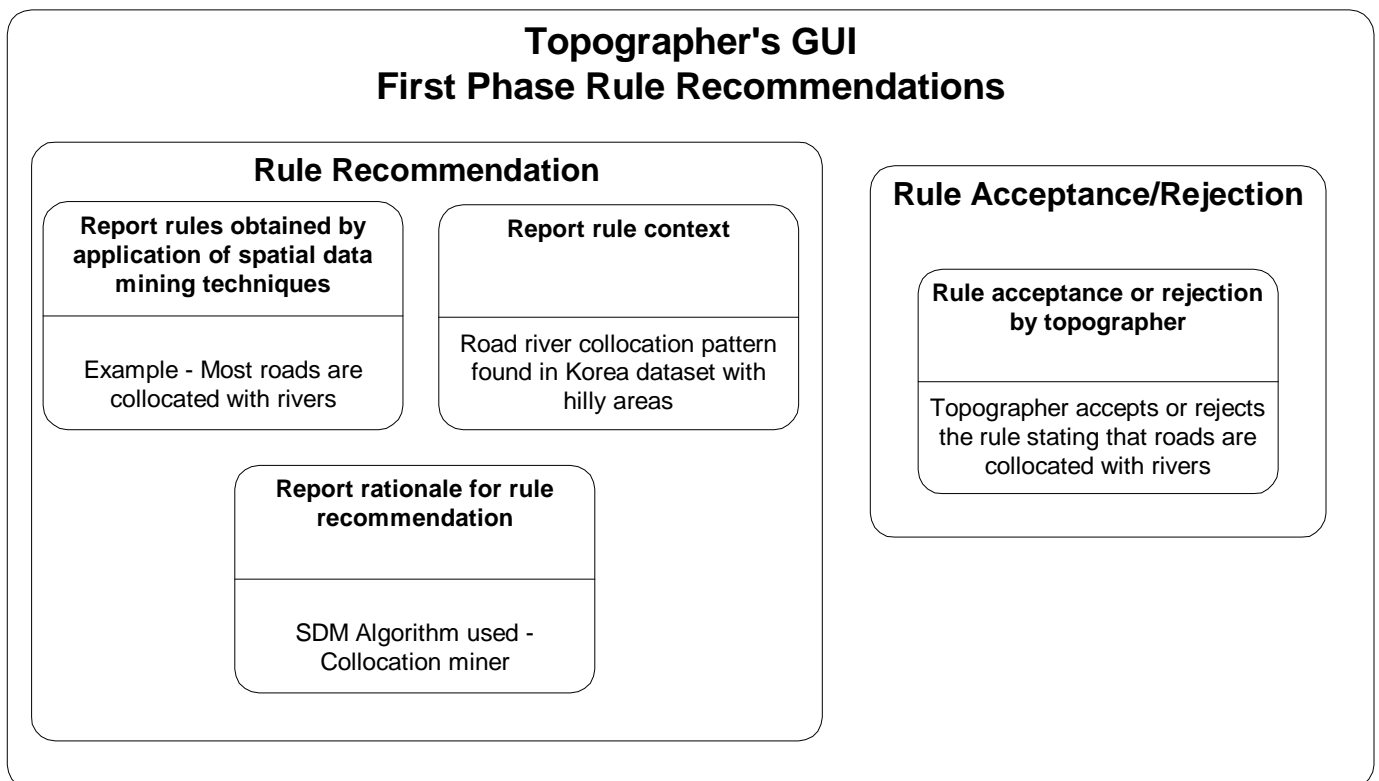
There are many non-trivial issues in merging multiple overlapping datasets about a common geographic region, particularly when they use different spatial frameworks. We do not intend to develop new techniques for such conversions. However, we would work with state of the art data merging tools and technologies in this area.

We will implement this feature on a best-effort basis, using the best practice in the area. For example, we will use GML, a popular XML standard for geographic data incorporating OGIS data types. This will allow TopoAssistant to import and export datasets with a wide variety of data file formats, which can be converted to GML. Similarly we will support ESRI shapefiles since ESRI provides translators between many common data formats and shapefiles. If a common MSDS data format cannot be translated to GML or shapefiles using COTS ontology mapping tools, we may develop translators for the selected custom MSDS formats. Our spatial dataset merging techniques can be used to build the **data conversion and import tool** depicted in the TopoAssistant conceptual architecture in figure 2.

### 2.5.4 Designing the Topographer’s GUI

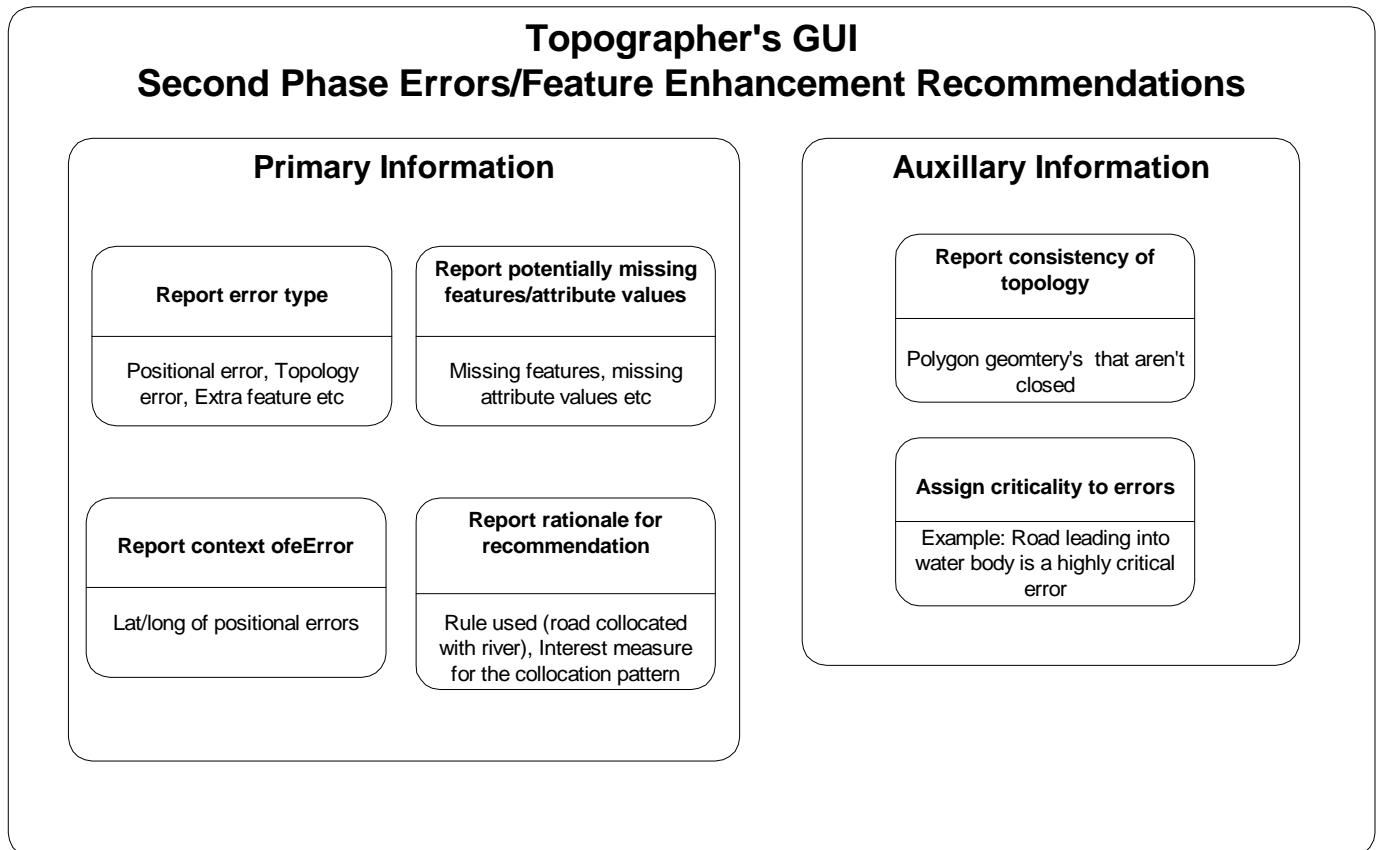
The topographer should be presented with actionable information from the spatial data mining process that can then be leveraged for verification/error detection and intensification/densification of the MSDS.

Information will be presented to the topographer in two phases. In the first phase (Figure 27) TopoAssistant’s spatial data mining techniques will be applied on the training data set to discover interesting rules. These discovered rules are then presented on the topographer’s GUI along with the context of the rule and the rationale for the recommended rule. The topographer is then given an option of accepting or rejecting the discovered rules. Domain experts and spatial data mining agent customizers will primarily use the information delivered on the GUI in the first phase. The domain expert topographer will then use this information to construct custom agents which are then added to the TopoAssistant system for error discovery and feature attribution.



**Figure 27: Topographer’s GUI – Report Discovered Rules to Domain Experts in First Phase**

In the second phase (Figure 28), the results of applying the discovered rules to the target dataset are presented to the topographers. Error types like positional error of road features and missing attribute values are predicted and then presented on the GUI. The error context and the rationale for the recommendation are also provided. Auxiliary information like topology consistency of the geometries and the criticality of the recommendation are also reported on the GUI. The topographer can either accept or reject these recommendations. The recommendations provided in the second phase will primarily be used by those topographers concerned with the final production phase of MSDS. To enable novice users to operate the TopoAssistant system, the topographer's GUI will be designed and implemented in an interview based GUI format.



**Figure 28: Topographer's GUI – Report Recommendations after Application of Valid Rules to Target Dataset in Second Phase**

### 3. Related Work

Architecture Technology Corporation (ATC) will leverage some of the pioneering research performed at the University of Minnesota in spatial data mining [<http://www.cs.umn.edu/research/shashi-group/sdm.html>] to develop the TopoAssistant. In addition we will also build on work done by other researchers in this area that were surveyed in [Koperski, Han, & Adhikari 1998] and [Shekhar et al 2004].

ATC has a proven track record of exceptional performance on research projects that produce innovative technologies. A representative sample of such projects is presented below.

**MobiWeb - Mobile Web and E-Mail Access over Satellite Links:** Developed for the U.S. Coast Guard under a DoT Phase II SBIR, MobiWeb is a new secure real-time messaging system for the agency that is designed to operate over commercial satellite services. It consists of three major components: a dial-on-demand router, proxy software, and a tunable TCP/IP stack. These components are used to provide secure and cost-optimized data communications between cutters and the shore-based command centers. MobiWeb has been successfully transitioned into the USCG operational environment. Four Coast Guard cutters were recently outfitted with this technology and plans are under way to refurbish another 70-75 cutters in the near future.

**Secure Multimedia over ATM Real-Time Services (SMARTS):** This Phase II SBIR effort for US Army CECOM built and demonstrated a prototype implementation of a set of network services called SMARTS that enables end-to-end security (i.e., authentication and confidentiality) of H.323 media transported over a bandwidth constrained ATM backbone network. Underlying the implementation of SMARTS are two key innovations: (1) an applique for based approach for H.323 media security that allows existing COTS H.323 products to be easily augmented with security using a “plug-in” software module; and (2) mechanisms for bandwidth optimized transmission of H.323 media over AAL5 and AAL2. The security applique for H.323, now called SecureIT!, is available as a licensable intellectual property (IP) offering from ATC. **This Phase II project is one of the six projects nationwide that has been selected for the 2002 Army SBIR Phase II Quality Award.**

**Datalets :** ATC is designing a data-oriented software entity that autonomously supplies data to requesting processes (fuselets) as part of the USAF Joint Battlespace Infosphere initiative. The concept of a datalet corresponds closely to that of the proposed ontology containing semantic information on the sources of sought-after data.

**Nomadic Routers for Rapidly Deployable Wireless Networks:** This DARPA Phase II SBIR effort successfully built full-scale prototypes of the nomadic router (wireless IP router) software for Windows and Linux platforms using 802.11 Wireless Ethernet adapters. Innovations include the Source-Initiated Ad Hoc Routing Algorithm (SARA) whose performance and capabilities support an asymmetric as well as symmetric ad hoc wireless network environments, a multicast routing protocol for ad hoc networks, and a cluster gateway node that interconnects an ad hoc cluster to a wired backbone network and the Internet. The wireless IP router products developed under this effort are currently offered as licensable software by ATC. The technology developed by this SBIR was also successfully transitioned to three DARPA BAA efforts.

**Techniques of Intrusion-Resistant Ad Hoc Routing Algorithms (TIARA):** This research effort, funded by DARPA/ATO’s Fault Tolerant Networks (FTN) program, developed a groundbreaking approach for protecting ad hoc networks against denial of service (DoS) attacks. ATC has successfully built and demonstrated a TRL 5 implementation of a TIARA based survivable ad hoc network. DARPA recognized this project with a “bytes-per-buck” award in January 2002.

**Randomized Failover Intrusion-Tolerant Systems (RFITS):** Sponsored by DARPA/IPTO’s Organically Assured Survivable Information Systems (OASIS) program, this research effort developed a collection of novel survivability design patterns for building DoS-resistant highly-available information systems. The project successfully built and demonstrated a TRL 6 implementation of a patent-pending software product called VPNshield for protecting VPN services against flooding DoS attacks. VPNshield is currently offered as a licensable IP offering by ATC. Other products implemented under the effort include FlowShield, a DoS-resistance mechanism for protecting specified IP packet flows from flooding DoS attacks, and JMSshield, a patent-pending middleware extension for protecting JMS applications from topic flooding DoS attacks. FlowShield technology was integrated within the Security Management System (SMS) being developed by CECOM.

## 4. Commercialization Strategy

### 4.1 Commercialization Strategy

ATC is committed to commercializing and marketing the products of this proposed new research and development effort. The company has a proven track record of taking ideas from the R&D stage to commercial products. After extensive in-house research in networking technology, ATC developed advanced prototypes of several networking products that were transitioned into its Triticom line of products. Furthermore, ATC has successfully transitioned products of several SBIR efforts into licensable intellectual property (IP) offerings. The following table summarizes some of these successes. Details on these IP offerings are available at [www.atcorp.com](http://www.atcorp.com).

Product	Technology Source	Comments
SecureIT!	CECOM Phase II SBIR	Installed in CECOM IVDV lab. Winner of 2002 Army SBIR Phase II Quality Award.
WrouteIT!	DARPA Phase II SBIR	Technology transitioned to Univ. of Minn. Distributed Robotics effort. Provided leverage for ATC's successful bid for an effort under DARPA's FTN program.
Scorecard/RISST	Navy/NSWC Phase II SBIR	Licensed by DoD vendor
VPNshield	DARPA OASIS program	Integrated within CECOM's SMS product and successfully demonstrated at CECOM
GeoTIDeS	Navy/ONR Phase II SBIR	Selected as a showcase technology by ONR's technology transfer support program
MobiWeb	DoT/Coast Guard Phase II SBIR	Currently operational on USCG cutters in Iraq. Planned induction into Coast Guard fleet by end of 2004.

ATC has the resources and capabilities to market the products of this research to the DoD as well as the commercial market. The company has a business development group headed by a Vice President to market its products, technologies, and services to the Government sector. ATC can market newly commercialized products through the same channels that we use for our Triticom brand of products. We recognize that the key to successfully marketing a product in the computer industry is a well-established distribution network for the product. Over the years, ATC has built a strong worldwide distribution network for its Triticom products that we will exploit to market the commercial products derived from TopoAssistant.

Our commercialization team will be led by Dr Ken Thurber who has over 20 years of experience in successfully commercializing R&D efforts into products.

#### **Kenneth J. Thurber – Commercialization Strategist**

Ph.D., Electrical Engineering, Montana State University, 1969.

#### **Experience**

Dr. Thurber is the President of Architecture Technology Corporation. He is a recognized expert in networking technology and computer architecture with over 25 years of R&D experience. He founded ATC in 1981. In 1990 he diversified its offerings to include a highly successful line of networking products that are marketed under the brand-name Triticom. He developed and

implemented the company's strategic plan for its entry into the networking products arena. Prior to founding ATC, he was with Sperry-Univac's Defense Systems Division, where he was the principal architect of several embedded computer systems (including the AN/UYK-43) that were successfully deployed in defense systems. While employed at Honeywell's Systems and Research Center, he led research projects that led to the development of special-purpose airborne and spaceborne multiprocessor systems. Dr. Thurber has authored over 60 papers and written 14 books in the areas of computer architecture and local area networks. He also served on the IEEE Computer Society Governing Board and as Chairman of the 6th, 7th and 8th Data Communications Symposia. He was nominated as a Golden Core member of the IEEE Computer Society. Dr. Thurber will serve as the productization strategist for this research.

## **4.2 Company Background**

Architecture Technology Corporation (ATC) was formed in 1981 as a consulting and publishing firm specializing in computer systems architecture. ATC is a privately held company headquartered in Minneapolis, Minnesota, with an office in Washington, D.C. and a subsidiary in New York. While providing services in a wide array of areas within the computer systems architecture field, ATC has specialized in computer networking technologies since the firm's inception. In 1990, it diversified its offerings to include commodity priced software and hardware products for computer networking. These products are sold under the brand name Triticom. In July 1999, ATC acquired Odyssey Research, now operating as ATC-NY. It is a wholly owned subsidiary specializing in Information Assurance. Currently, the combined organizations have about 85 employees. Corporate operations include: Engineering Services, Triticom (Commercial Products), ATC-NY, and corporate Research and Development. Early on, the company recognized the need to employ a management paradigm for its commercial product development effort that was different from the rest of its operations. Therefore, the product development group was set up as a semi-autonomous operation within the company's headquarters in Minnesota, borrowing the brand name of its products, Triticom. The following paragraphs provide an overview of each of the operations of the company.

### **Engineering Services**

ATC provides systems engineering solutions to a number of key government agencies and prime contractors. Our expertise has centered around sophisticated local area network-based, fault-tolerant, distributed computer architectures. We have provided leadership and technical support in the areas of: Risk Assessment/Reduction, Monitoring & Control, System Test, WAN Communications, Interface Design, Internetworking, Hardware & Software Prototype Development, and several specialty technology areas. Our list of satisfied, repeat customers includes: IBM, Cray Research, 3M/Interactive Systems, TRW, Smithsonian Institution, U.S. FAA, DOT, and EPA, Educational Broadcasting Corporation, Automated Controls Incorporated, BankAmerica, Merrill-Lynch, Wolffer's, AT&T Information Systems, University of Pennsylvania, University of Minnesota, Digital Equipment Corporation, ALCOA, Ford Motor Company, Hughes Aircraft Company, and United Defense LP.

### **Triticom (Commercial Products)**

Since 1990, ATC has been developing a set of private-label software products for computer networking that is marketed and distributed under the trade name "Triticom." Triticom products have earned a number of awards for quality including the LAN Magazine Product of the Year (Protocol Analyzer Category), Data Communications Magazine Tester's Choice Award, Network World Shortlist, and two InfoWorld Quality Assurance awards. Triticom's latest offering is the **LANdecoder Network Management Suite**. LANdecoder NMS incorporates all the elements needed for a comprehensive network management tool into one integrated product offering. The foundation of LANdecoder NMS is Triticom's award winning network analysis product,

LANdecoder32 V3.0. LANdecoder32 provides the platform to monitor network operations and health, and capture selected data to aid in diagnosing and isolating network problems.

#### **ATC-NY, Inc.**

Located in Ithaca, New York, ATC-NY has been providing products, as well as consulting, in the fields of computer information security and tool development for high-assurance software since 1983. ATC-NY has over 15 staff members, most with advanced degrees in mathematics and computer science, specializing in Information Security. Their products provide protection to organizations, both government and commercial, from unauthorized access and malicious attack. In addition, Safety-Critical Systems provide solutions to software failures, saving time, money and possibly lives. Clients include multinational companies like General Electric, Microsoft, and Boeing Aerospace and such well-known government agencies as Defense Advanced Research Projects Agency (DARPA), NASA, U.S. Army's Tank and Automotive Command (TACOM), NLM, and Air Force Research Lab (AFRL). ATC-NY's research in Information Security has resulted in products, architectures, and policies addressing the protection of sensitive information, the detection of system intrusions, and the analysis of system vulnerabilities. These areas of work include: **Intrusion and Fraud Detection, Computer Forensics, Internet/Intranet Security, ATM Network Security, Secure Collaboration and Workflow/Healthcare Informatics/Telemedicine, Object Based Security Solutions, Security Policy Analysis and Design and Vulnerability Analysis.** In Safety-Critical Systems, ATC-NY's research has focused on the application of formal mathematical techniques to the analysis and verification of safety-critical software and hardware.

#### **Research and Development**

The mission of the corporate R&D function is to perform high-quality, high-risk, high-payoff engineering research in cutting-edge computer technologies and information assurance that could produce significant benefits to ATC's business in the areas of product development and consulting services. Corporate R&D efforts focus on the following technology areas that fall under the broad umbrella of distributed computing systems: **Next-Generation Network Services, Network Management and Test Tools, Survivable Information Systems, Intelligent Systems.** In 1994, R&D efforts expanded to include participation in the Small Business Innovative Research (SBIR) program. Since then, the organization has received multiple SBIR awards from a wide range of sponsors including: Defense Advanced Research Projects Agency (DARPA), National Science Foundation (NSF), Department of Defense (DoD), NASA, Department of Commerce (DoC), Department of Energy (DoE), and the Department of Education (DoEd), covering the technology focus areas listed above. R&D's efforts have resulted in a number of Phase II awards that are progressing into product development. This expertise has been critical in teaming with other companies on successful contract awards managed by DARPA and other government agencies. *In 1998 and 2000, the Small Business Administration presented ATC the prestigious Tibbetts Award for outstanding performance in the SBIR Program. In 2002, ATC received the US Army SBIR Quality Award.* As part of our technology marketing strategy, we are forging strategic relationships with a number of Fortune 500 companies including: Alliant TechSystems, GTE/BBN, TRW, United Defense, Boeing, General Dynamics Information Systems, Rockwell Science Center, Honeywell Technology Center, Lockheed Martin, Raytheon and others.

## **5. Conclusion**

The TopoAssistant tool developed by this SBIR effort will significantly reduce the time and effort expended by Army topographers in generating MSDS to support operational mission requirements. It will enable them to address the responsiveness requirements for the Objective Force. This tool directly addresses the needs for applications such as Intelligence Preparation for

the Battlefield (IPB), tactical decision aids (TDAs), corridor analysis, and route planning within the Future Combat Systems.

To establish technical feasibility during Phase I, we focused on evaluating concept feasibility and implementation feasibility of the TopoAssistant approach. Concept feasibility was evaluated by using a benchmark dataset from TEC with half a dozen layers and numerous features describing a region in Korea. Spatial data mining techniques were able to identify interesting, useful and non-trivial patterns relating to MSDS verification and densification. These patterns were reviewed by experts in TEC for usefulness in the MSDS refinement process by comparing the discovered patterns against auxiliary map layers. The effectiveness and performance of the TopoAssistant tool for map error detection and feature enhancement exceeded expectations during the Phase I effort. We identified the technical challenges that need to be addressed when implementing a full-scale operation prototype of the TopoAssistant. We also formulated our technical approach for identifying each of these technical challenges during the follow-on Phase II effort.

## 6. References

- [Agrawal & Srikant1994] Agrawal, R., and Srikant, R. 1994. Fast algorithms for Mining Association Rules. In Proc. of Very Large Databases.
- [Barnett & Lewis1994] Barnett, V., and Lewis, T. 1994. Outliers in Statistical Data, John Wiley, 3rd edition.
- [Cressie1993] Cressie, N. 1993. Statistics for Spatial Data (Revised Edition), New York: Wiley.
- [Han, Kamber, & Tung2001] Han, J.; Kamber, M.; and Tung, A. 2001. Spatial Clustering Methods in Data Mining: A Survey. In Miller, H., and Han, J., eds., Geographic Data Mining and Knowledge Discovery. Taylor and Francis.
- [Hawkins1980] Hawkins, D. 1980. Identification of Outliers, Chapman and Hall.
- [Kabinier TEC] Debra Kabinier, "Digital Topographic Data (DTOP): Framework of Mission Specific Data Sets (MSDS)", [www.amso.army.mil/terrain/library/dtop.doc](http://www.amso.army.mil/terrain/library/dtop.doc).
- [Koperski & Han1995] Koperski, K., and Han, J. 1995. Discovery of Spatial Association Rules in Geographic Information Databases. In Proc. Fourth International Symposium on Large Spatial Databases, Maine. 47-66.
- [Koperski, Han & Adhikari 1998] K. Koperski, J. Han, and J. Adhikary, "Mining Knowledge in Geographical Data, *Communications of ACM*, 1998
- [Morimoto2001] Morimoto, Y. 2001. Mining Frequent Neighboring Class Sets in Spatial Databases. In Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
- [Shekhar et al.2002] Shekhar, S.; Schrater, P. R.; Vatsavai, R. R.; Wu, W.; and Chawla, S. 2002. Spatial Contextual Classification and Prediction Models for Mining Geospatial Data. *IEEE Transaction on Multimedia* 4(2).
- [Shekhar, Lu, & Zhang2001] Shekhar, S.; Lu, C.; and Zhang, P. 2001. Graph-based Outlier Detection : Algorithms and Applications (A Summary of Results). In Proc. of the Seventh ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 2001.
- [Shekhar & Chawla2002] Shekhar, S., and Chawla, S. 2002. A Tour of Spatial Databases. Prentice Hall (ISBN 0-7484-0064-6).
- [Shekhar & Huang2001] Shekhar, S., and Huang, Y. 2001. Co-location Rules Mining: A Summary of Results. Proc. of Spatio-temporal Symposium on Databases.
- [Shekhar et al 2004] Shekhar, S., Zhang, P., Huang, Y., Vatsavai, R., "Spatial Data Mining," as a book chapter to appear in "Data Mining: Next Generation Challenges and Future Directions", Hillol Kargupta and Anupam Joshi (editors), AAAI/MIT Press, 2004 (draft copy of chapter at <http://www.cs.umn.edu/research/shashi-group/sdm.html>)
- [Witten & Eibe 1999] Ian H. Witten, Eibe Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, October 1999.

[Shekhar & Lu2003] A Unified Approach to Spatial Outliers Detection (with C. Lu and P. Zhang), GeoInformatica: An Intl Jr. on Adv. of Computer Sc. for Geographic Info. Systems, 7(2), 2003 (A summary appeared in the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining, 2001).

[State Dept 1999] <http://www.usconsulate.org/hk/uscn/state/1999/0706.htm>.

[Army 2003] <http://www.army.mil/features/507thMaintCmpy/AttackOnThe507MaintCmpy.pdf>.

[SQL3/OGIS 2004] <http://opengis.org/docs/99-049.pdf>

[PostGIS 2004] <http://postgis.refrations.net/>

[Postgres 2004] <http://www.postgresql.org/>

[GEOS 2004] <http://geos.refrations.net/>

[QGIS 2004] <http://qgis.sourceforge.net/>