Fall 11-28-2011

Efficient Traffic Crash and Snow Complaint GIS System

Anthony B. Ngo
University of Nebraska-Lincoln, Anthony67.Ngo@Gmail.com

EFFICIENT TRAFFIC CRASH AND SNOW COMPLAINT GIS SYSTEM

by

ANTHONY B. NGO

A THESIS

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfillment of Requirements

For the Degree of Master of Science

Major: Computer Science

Under the Supervision of Professor Peter Revesz

Lincoln, Nebraska

November, 2011

EFFICIENT TRAFFIC CRASH AND SNOW COMPLAINT GIS SYSTEM

Anthony B. Ngo, M.S.

University of Nebraska, 2011

Adviser: Peter Revesz

We describe the design and implementation of a traffic crash and snow complaint GIS system developed for the Lincoln Public Works department. We also describe a novel geocoding algorithm that was used to move data from the older Criminal Justice Information System, which is a relational database, to the new GIS system. In addition, we describe the implementation of several indexing algorithms that enable the system to efficiently answer rectangular range queries and queries about the relative locations of moving objects. Finally, in many applications (on-line analysis or mobile GIS), we need to execute spatial query efficiently (fast and small), and to scan through all the existing complex objects (non-zero extent) might be very slow. Thus, we introduce an approximated indexing method, called ApproximatedR-Tree, for improving performance and space over complex objects.

Acknowledgements

I would like to express my sincere gratitude to my thesis advisor Professor Peter Revesz, who first introduced me to spatial databases and GIS. Without his expertise and generous guidance I would not have been able to complete this thesis. He has provided me numerous ideas for carrying out the research involved in this work.

Also, I would like to thank Dr. Jitender Deogun, and Dr. Ashok Samal for serving on my thesis committee. I appreciate their feedback on my research and their valuable comments on my thesis.

In addition, I would like to thank Alicea McCluskey, Shane Dostal, Scott Opfer, Al McCracken, and Angela Chesnut of the Lincoln Public Works Department for providing information and suggestion about street maintenance operations and traffic crash analysis; Mark Wieting of Lincoln Information System Department for providing traffic crash data from Criminal Justice Information Systems; Frank Larson, Tracy Schuppan, and Tan Pham of Lincoln Comprehensive Engineering Information Services (CEIS) Department for providing GIS technical advice; James Anderson of Lincoln Information Systems Department for supporting the ArcGIS Server; Virendra Singh of Lincoln Public Works Department for expert advice regarding traffic crash analysis; Tim Pratt of Lincoln CEIS Department for managing and overseeing software development, and Roger Figard for allowing and encouraging me to publish these results.

Moreover, I would like to thank Michael Boscarino, HP CoSD Applications Technical Lead, Harish Raju, HP CoSD Applications Portfolio Manager, Ross Martin, County of San Diego GIS Manager, and Jason Batchelor, County of San Diego GIS Coordinator, for encouraging and allowing me to work from home in Lincoln, NE in order to complete my Master thesis.

Finally, this thesis is for my beloved wife, Thuloan Pham, my daughters, Mary Ngo, Kristy Ngo, and Julie Ngo, and my parents. Without their day-to-day support, love, and encouragement, this thesis could not have been completed.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1

Introduction

City governments recently have provided more GIS-based public services. These systems allow city officials and the public to access to real and accurate data and planning tools which were not previously available [14]. In this thesis, we describe our design and development of the Lincoln Public Works (LPW) department's traffic crash data, and snow complaint GIS system. In addition, we describe the implementation of several indexing algorithms that enable the system to efficiently answer spatial queries and of the Approximated R-tree.

1.1 Motivation for traffic crash database

The Lincoln Public Works and the Lincoln Police Department generated and maintained for many years the traffic crash reports and associated statistics using the Criminal Justice Information System (CJIS), which is a relational database system. In 2008, the CJIS system made a recording of 7,890 crashes with a total financial loss approximately about 200 million dollars including wage loss, medical expenses, property damage, and insurance administrative costs [9].

On October 2009, LPW decided to convert the CJIS data to a new GIS-based system for mapping and analyzing traffic crashes in order to conduct a detailed crash study throughout the city of Lincoln. The primary goal of this project, called *CJIS to GIS*, was to provide a high performance GIS-based tool for identifying frequent traffic crash locations and provide realistic solutions to reduce the total number of traffic crashes and their monetary impact [9]. Identifying frequent or otherwise problematic traffic crash locations will lead to a better prioritization and planning of transportation improvements and to developing more effective countermeasures that address the identified crash patterns at all locations having frequent traffic crashes.

1.2 Motivation for snow complaint database

According to a New York Times article [30] about New York City government and a large snow storm:

> "The blizzard prompted a political crisis that became legendary in the annals of municipal politics, nearly brought down the administration of Mayor John V. Lindsay and offered an instructive lesson to elected officials in the politics of snow removal."

To avoid such a political crisis and other damages, cities need to be prepared for winter snow storms and have a plan to keep the streets cleared and safe. To manage snow removal efforts, cities need GIS-based systems that enable real-time visualization of snow-related complaints, such as "street is icy", "parking ban info", "snow push into street", "property damage", "plowing wrong side street," etc.

We combine crash reports data and snow complaints data in the same database because they are closely related. According to city statistics, January is a high-fatality month with many traffic crashes due to snow-related issues. A common GIS-based system for managing traffic crash and snow removal data can help visualize and analyze this relationship more deeply. That would help improve street improvement operations, which would reduce the number of snow-related traffic crashes.

1.3 Outline of the thesis

This thesis is organized as follows. Chapter 2 describes related work. Chapter 3 presents a novel and efficient geocoding algorithm that converts CJIS traffic crash data into GIS data. Chapter 4 describes an implementation of rectangular range queries. Chapter 5 gives an overview of the snow complaints database. Chapter 6 describes the implementation of an efficient moving points estimation algorithm. Chapter 7 describes the efficient approximated indexing algorithm, called an approximated R-Tree, which is based on the R-Tree. Finally, Chapter 8 suggests some future work.

1.4 The major contributions of the thesis

1. Development of an algorithm that efficiently converts relational traffic crash data into GIS spatial data. The details of this algorithm are described in Chapters 3 and 4 of this thesis and in Ngo and Revesz [19].

2. Development of an algorithm that efficiently translates relational snow complaints inputs into GIS spatial data. The details of this algorithm are described in Chapters 5 of this thesis and in Ngo and Revesz [19].

3. Development of algorithms to efficiently answer range queries on moving points. The detail of this work is described in Chapter 6.

4. Development of an approximate R-tree method to increase the efficiency of retrieval of spatial data. This algorithm and its implementation for a mobile fire damage assessment application are described in Chapter 7.

Chapter 2

Related Work

1.1 Crash Data Analysis

Some recent research considered mapping only intersection crashes using GIS technology. Nwaneri [20] describes the application of GIS to map intersection crashes at only 20 intersections for 24 months of study. Parrish, Dixon, et al. [22] discuss an efficient tool for transportation safety engineers and policymakers to analyze the traffic crash data typically obtained from police reports without using GIS technology. Bao, Ying, et al. [2] reports their development of a GIS-based traffic network analysis system, which provides a graphical analysis platform to transportation planners and researchers for transportation network analysis. However, almost all of these known public researches discussed to date fall into one of three categories. Either they are just a non-GIS crash data analysis tool, or they are a GIS crash data analysis tool, or they are so simple that they are not useful in solving many traffic crash problems that involve geographic visualization and querying.

2.2. Geocoding

Other existing traffic crash databases such as [7], [31], and [32] use traditional geocoding, which requires postal addresses. However, we cannot apply for our application because LPD never has recorded any postal address. On the other hand, Zhang, Chen, at al. [36] introduce a new type of geocoding task as segment-type geocoding in contrast with classic geocoding that takes a street address or intersection. This approach is so specific that it is useful in solving only one or two problems. Moreover, this approach is dependent on different types of street intersections (one node or multiple nodes). In this paper we describe a *not-too-general* or *not-too-specific* paradigm that can be precisely specified and can be used to solve many traffic crash related problems.

Our traffic crash and snow removal system has unique indexing features. As a result, our system is able to answer efficiently many queries of the type “find the number of (traffic crash or snow complaint) events on the map on a particular street segment.” Efficient answering of these types of queries is based on the implementation of the main indexing algorithm in [23], which in turn is a based on Bentley’s ECDF-trees [5].

ESRI’s ArcGIS system provides some visualization tools for spatial analysis. However, those tools are limited to small amount of data because of inefficient processing methods implemented within ArcGIS. For example, the ArcGIS query task default parameter *maxresultsize* is normally set to 500 (version 9.x) and 1000 (version 10.x). Our system avoids this limitation and could handle much larger data sets by using the indexing algorithm in [23, 24].

1.3 R-trees

The R-tree introduced by Guttman [12] is one of the most popular access methods for rectangular spatial objects. The R-tree is a height-balanced tree with index records in its leaf nodes containing pointers to rectangular spatial data objects. We can also use an R-tree index to deal with non-rectangular objects by drawing around each non-rectangular object a *bounding box*, which is a rectilinear shape (rectangles and hyper-rectangles) that completely contains the bounded object but is the smallest possible rectilinear shape.

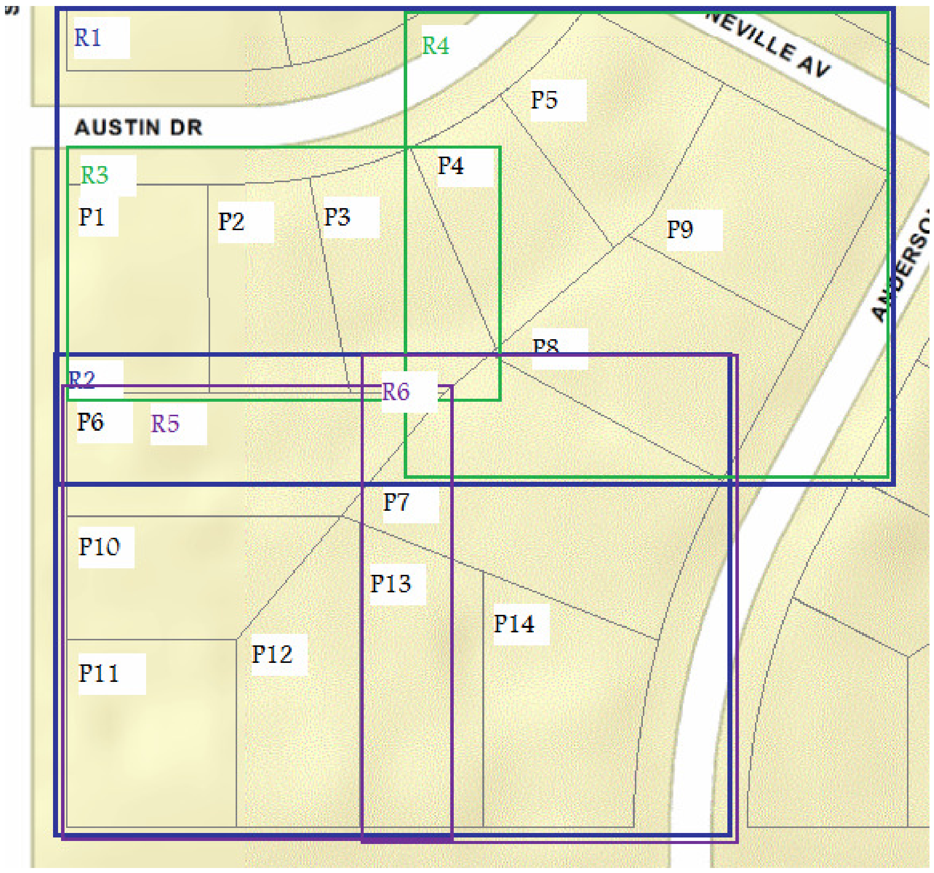


Figure 1: R-Tree Bounding Boxes

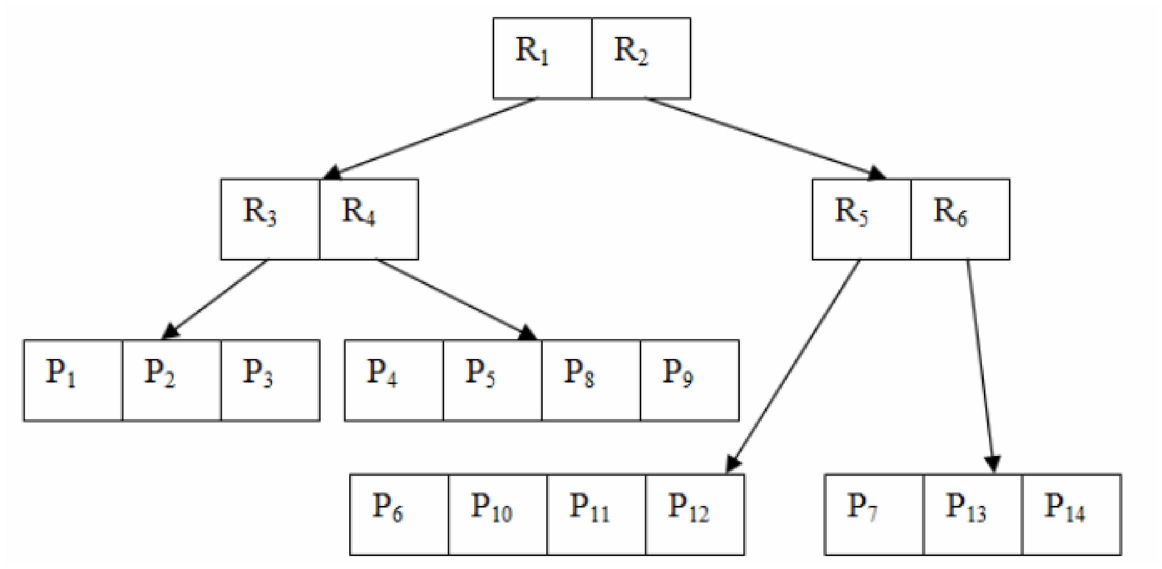


Figure 2: R-Tree Index Structure

Example 3.1 Figure 1 shows a piece of a city map with fourteen land parcels marked as $P_1 \ldots P_{14}$. Each of the fourteen land parcels can be associated with a bounding box, which is the smallest enclosing rectangle with sides parallel to the x- and the y- axes. These fourteen bounding boxes form the leaves of the R-tree shown in Figure 2. Subsets of the bounding boxes at the leaves can be grouped together into larger bounding boxes associated with the internal nodes of the R-tree. For example, Figures 1 and 2 show that the bounding boxes for the land parcels P_1, P_2 and P_3 can be grouped together into the R-tree internal node R_3.

In an ideal R-tree, we can search for the location of any point from the root to a single leaf. However, R-trees may not always partition the space into disjoint cells. Therefore, the point we are searching may belong to the bounding boxes of several internal nodes. In these cases, the search must branch and continue in each of the sub-trees rooted at those internal nodes. Hence in the worst case scenario, where we repeatedly branch to all possible sub-trees, the computational time of the search will be linear in the size of the database the entire database. This well-known limitation is a drawback for using R-trees where the response time is critical. Several works, including [6, 8, 11, 13, 16, 21], give proposals to improve R-trees.

Chapter 3

The Geocoding Problem

Geocoding is the problem of converting non-geographical information, such as street address, into valid geographical information that GIS systems or constraint database systems [15] can use, typically (x, y) coordinates. For example, the CJIS relational database records *traffic crash report data* in a form which needs to be converted to a GIS form. Since CJIS is a very large database, manual conversion of the data is prohibitively expensive. It may take between 15 and 30 minutes for a GIS specialist to manually add one *traffic crash report* into an ESRI ArcMap/ArcInfo shapefile. Since CJIS has over 100 thousand traffic crash report data, it may take between 25 and 50 thousand hours to manually convert each traffic crash report data to a GIS data. Hence we developed an efficient conversion algorithm that automatically converts the old data into the GIS representation. Now let us analyze the problem of locating a traffic crash on a GIS map. There are two cases, either the crash happens at an intersection or at a non-intersection location, which is at the middle of a segment of a street. The first case is easy and can be converted by an ESRI ArcGIS geocoding tool. The second case is not solved by the ArcGIS tool. Hence we focus on the latter case. In contrast to the segment-type geocoding method in [36], our geocoding method is independent of the different types of street intersections because we use the street center line GIS map shown in Figure 3, which yields only a single intersection node for streets that cross each other.

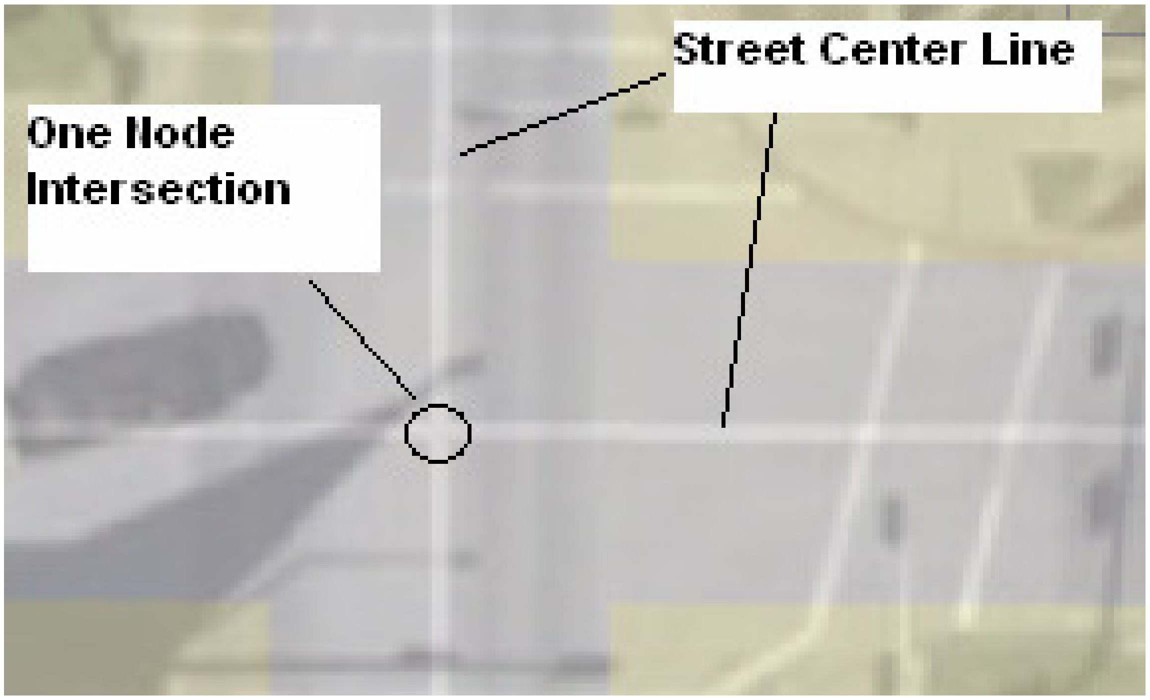


Figure 3: Street center line and one-node intersection.

Each CJIS traffic crash report is purely relational and does not include any information about location coordinates (latitude and longitude) or street address used in traditional geocoding [3]. Instead, as shown in Table 1, the CJIS database gives the names of *OnStreet, AtStreet*, and *BtStreet*.

The particular geocoding problem that we considered is to find the GIS street segment, called *midBlockStreet*, on *OnStreet*, where the traffic crash happened. Before presenting a solution, let us first recall the following important definitions regarding point dominance [23]:

Definition 1. Point $Q(x_Q,y_Q)$ dominates point $C(x_C,y_C)$, written as $Q > C$, if and only if $x_Q > x_C$ and $y_Q > y_C$

Definition 2. A block street dominates point $C(x_C,y_C)$, written as blockStreet > C, if and only if for every point $Q(x_Q,y_Q)$ on blockStreet $x_Q > x_C$ and $y_Q > y_C$

Table 1. Example CJIS traffic crash street information

ONST	**ATST**	**BTST**
P ST	N 11TH ST	N 12TH ST
N 27TH ST	Vine ST	T ST

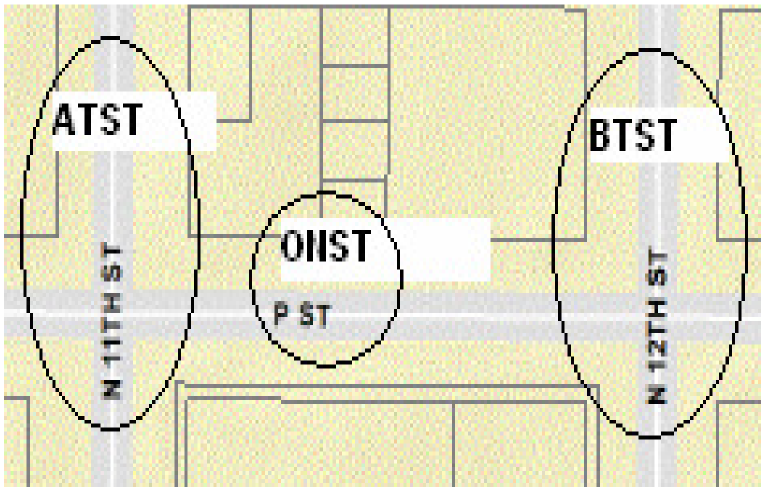


Figure 4: Searching street segment on ONST between ATST and BTST.

Now we can give the following solution:

Let I_1 be the intersection point at ONST and ATST.

Let I_2 be the intersection point at ONST and BTST.

Find *midBlockStreet* on ONST between ATST and BTST.

Without loss of generality, we assume that $I_1 < I_2$. Then the following must hold:

$$I_1 < \textit{midBlockStreet} < I_2 \tag{1}$$

Below are two steps to find midBlockStreet.

Step 1. Find *onStreetSegmentsSe*t, which is the set of all street segments of ONST.

Step 2. Find *midBlockStreet*, which is the street segment in *onStreetSegmentsSet* that is dominated by I_2, and dominates I_1, by condition (1).

In the GIS database the middle point of *midBlockStreet* can represent the approximate traffic crash location.

Chapter 4

Rectangular Range Queries

For developing a practical crash analysis tool, we need to provide a function that finds all crashes that happened in a rectangular area. Such a function could be used to identify problematic rectangular areas with a high frequency of traffic crashes.

Consider the red rectangular area shown in Figure 5. The location of a traffic crash C(x,y) lies within the rectangle if C(x,y) is dominated by the upper right point Q(x,y) and C(x,y) dominates the bottom left corner point P(x,y). Hence we have:

$$X_C < X_Q \text{ and } Y_C < Y_Q \text{ and } X_P < X_C \text{ and } Y_P < Y_C$$

To implement dominance queries efficiently, we store all traffic crash records in an ECDF-tree, which runs in O(log N) time for each dominance query and requires only O(N logN) space where N is the number of traffic crash records. We investigated the computational complexity of finding traffic crashes dominated by a point Q. We experimentally compared the runtime performance of two algorithms for finding traffic crashes dominated by a point Q; the first algorithm used ECDF-trees [5] and the second is algorithm did not use any indexing.

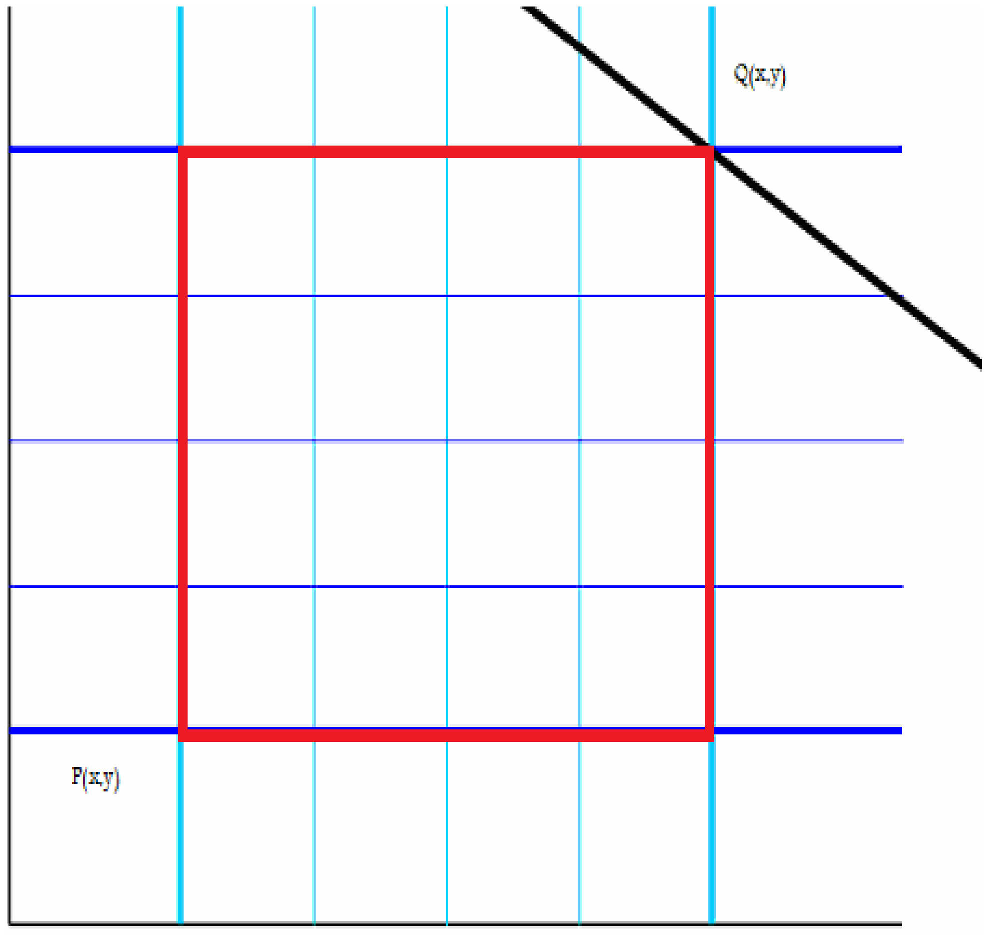


Figure 5: Traffic crashes in the red rectangular area are dominated by point Q(x,y) and dominate point P(x,y).

The results were visualized by the system. For example, Figure 6 shows traffic crashes dominated by point Q for January 2010.

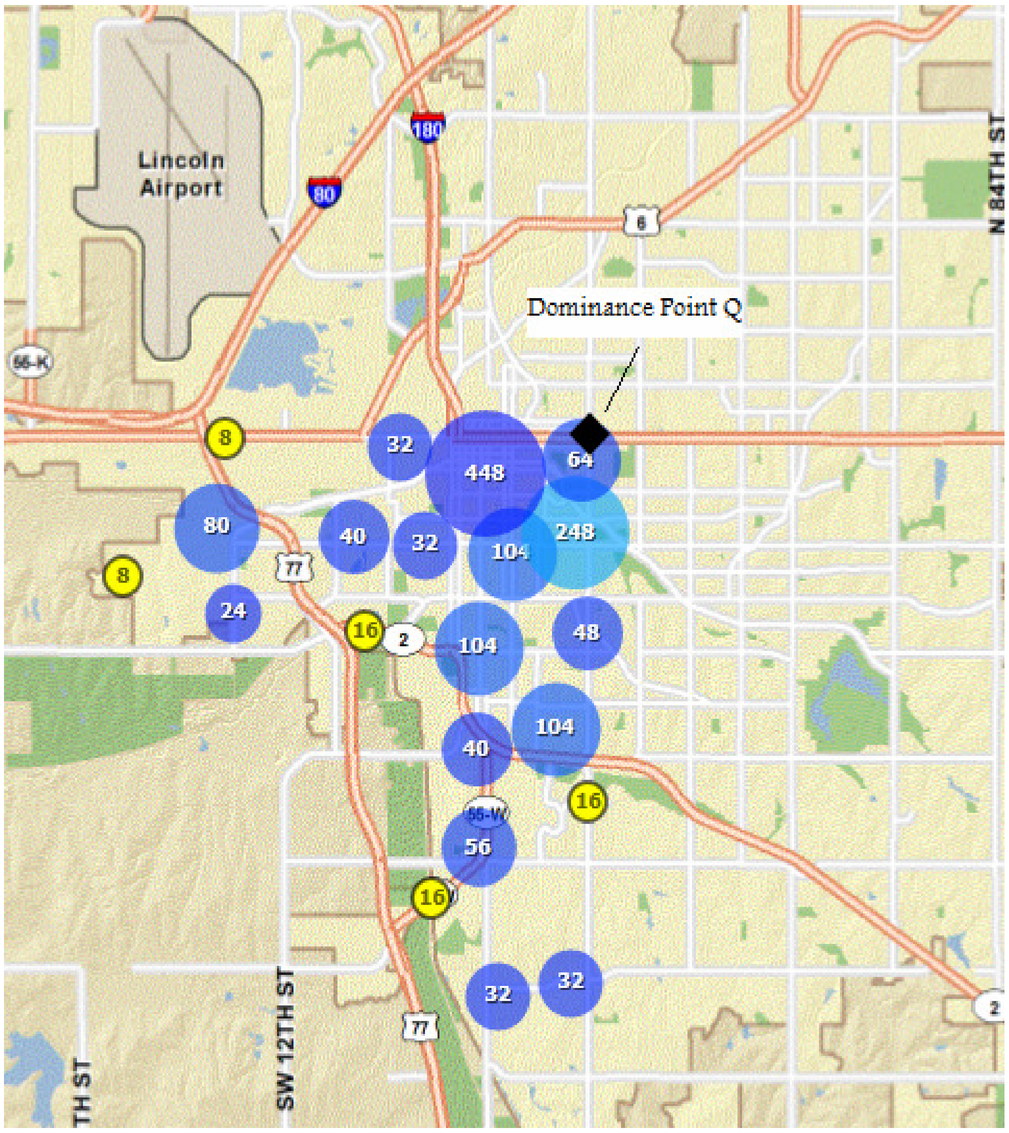


Figure 6: Traffic crashes dominated by point Q for January 2010.

Figure 7 and Table 2 show that as the total number of traffic crashes increase, the implementation with indexing is running faster than the one without indexing.

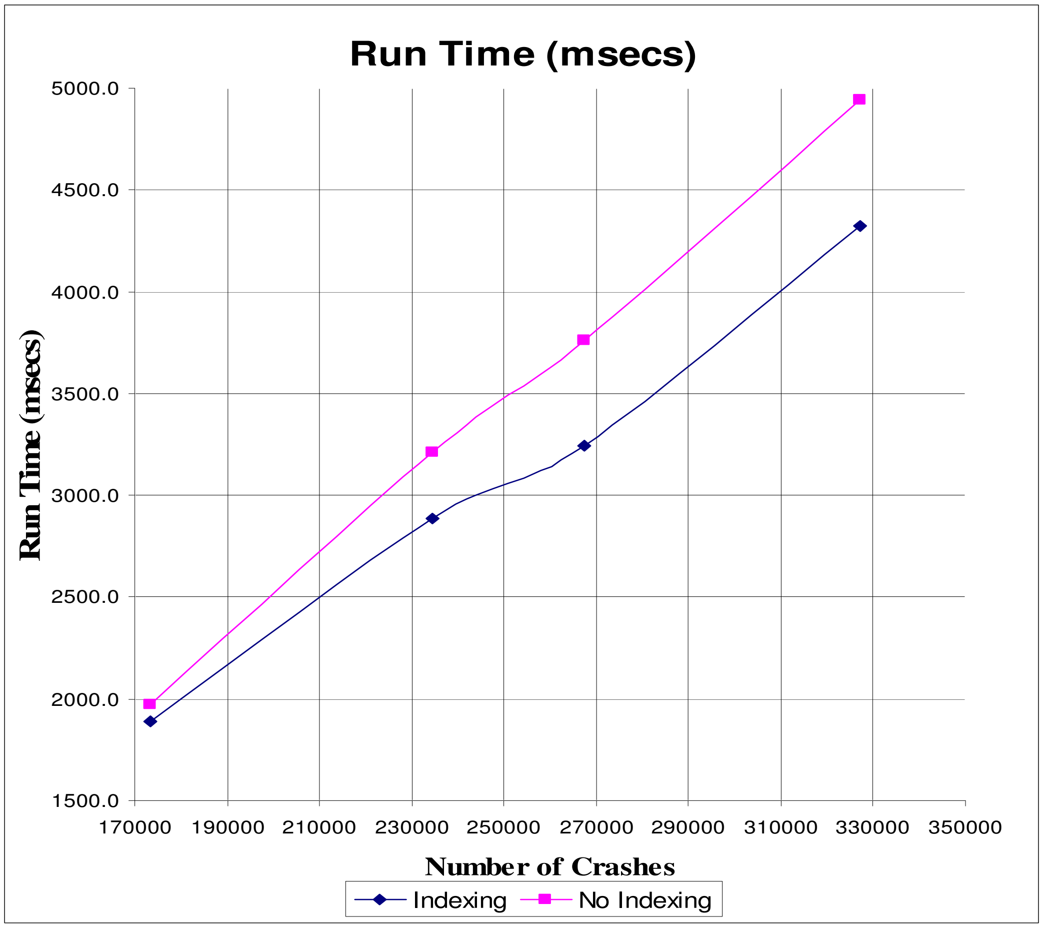


Figure 7: Run time using indexing versus no indexing.

Table 2: Run times using indexing versus no indexing.

Number of Crashes	**Dominated Points**	**Indexing Run Time (msecs)**	**No Indexing Run Time (msecs)**
7358	2029	46.8	46.8
14899	4134	73.0	78.0
173200	48028	1887.6	1971.0
234580	65448	2888.0	3213.6
267496	74856	3244.8	3760.0
327092	91404	4321.2	4945.2

Chapter 5

The Snow Complaint Database

The city of Lincoln is prepared in case of any snowstorm to implement its snow removal plan. There are currently three snow removal districts, namely, the North Eastern, the South Eastern, and the Western districts. In the past, street maintenance operations were maintained using Microsoft Access and Microsoft Excel applications for recording labor, snow removal equipments, material, and snow operations. The Access and Excel methods are inefficient for managing snow removal operations as the city grows. In addition, in order to create a GIS map showing the current snow complaints, only a professional GIS specialist knows how to manually geocode a snow complaint. That poses a significant challenge to less-skilled end users who may need to use GIS technology. Hence there was a need to have a better tool for data management, especially applying GIS technology by less-skilled end users to handle day-to-day street operations. That was a major motivation for developing the *Street Maintenance Manager* application shown in Figure 9. The *Street Maintenance Manager* provides a Lincoln snow complaint map (Figure 11) that has been designed and developed using the current cutting-edge ESRI ArcGIS technology and state of the art non-traditional geocoding for non-GIS end-users.

The snow complaint application also has the same two main problems as the traffic crash application, that is, geocoding and rectangular range querying, which have already been addressed above. Therefore, we will not repeat them in this section. Instead we would like to present in Figure 8 an overview of the Street Operation Manager system.

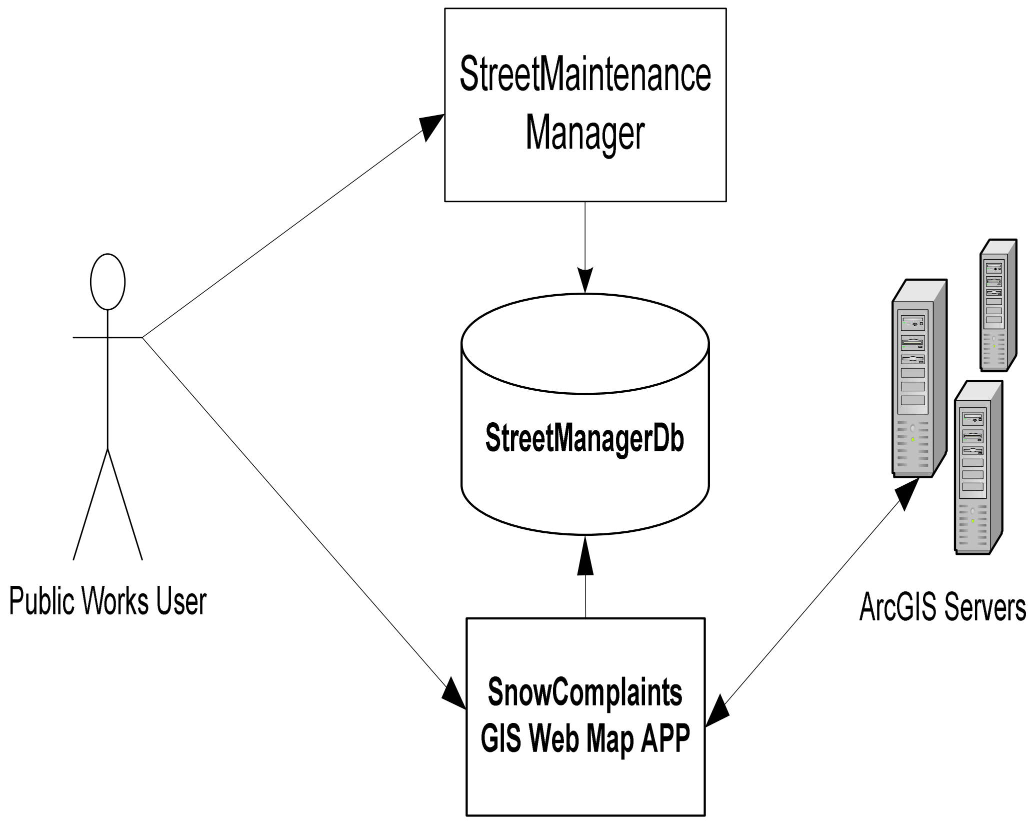
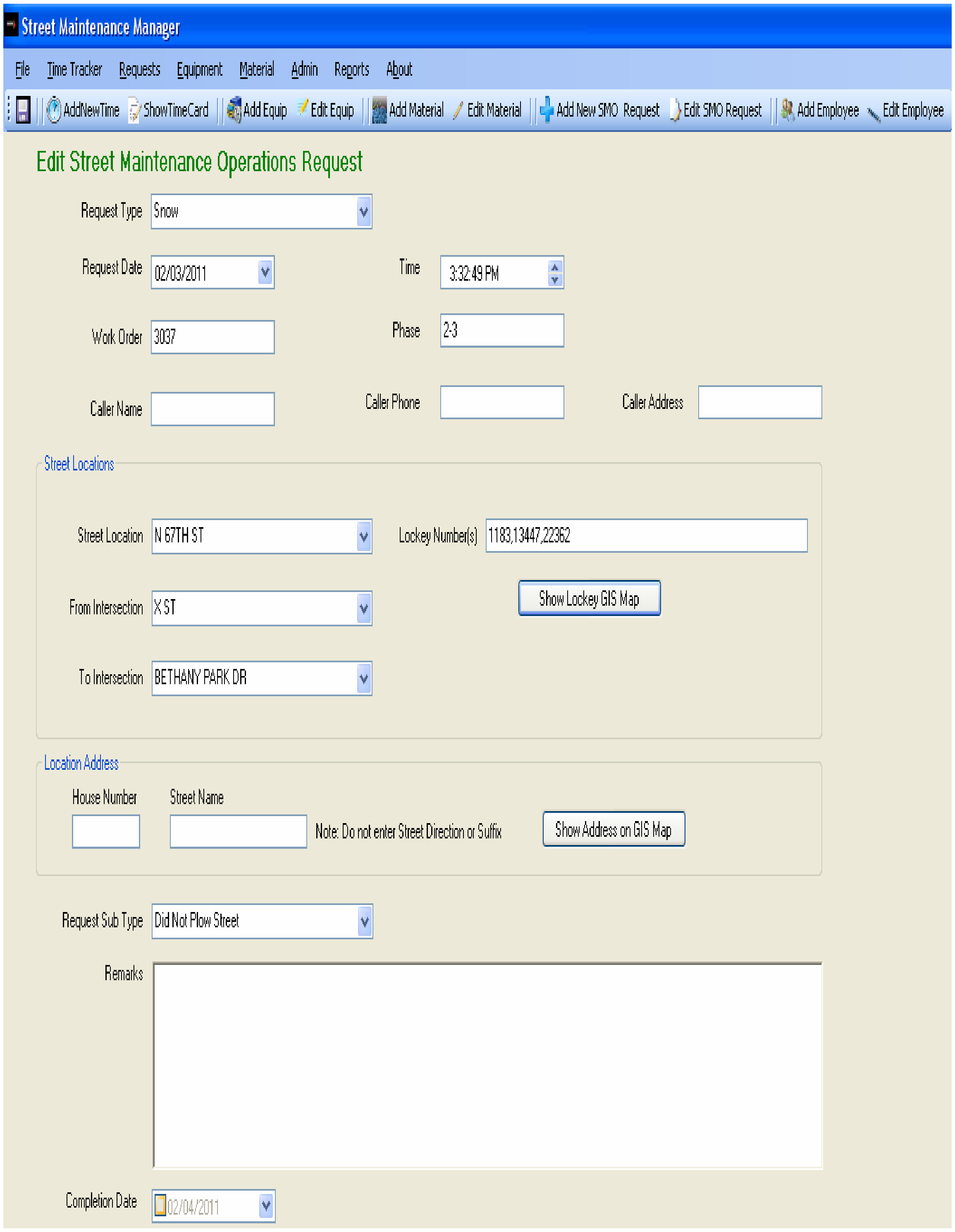


Figure 8: Street Maintenance Manager Software architecture.

We developed the *Street Maintenance Manager* using the following steps. First, we developed a day-to-day input form including some street location fields for geocoding, such as, On Street, At Street, and Between Street, for recording a snow complaint as shown in Figure 9. The input form also has a built-in street dictionary to quickly correct a street name, making sure that the complaint location does exist. This feature ensures that snow crews are sent to the correct locations.


Street Maintenance Manager
File Time Tracker Requests Equipment Material Admin Reports About
AddNewTime ShowTimeCard Add Equip Edit Equip Add Material Edit Material Add New SMO Request Edit SMO Request Add Employee Edit Employee
Show Lockey GIS Map
Show Address on GIS Map

Edit Street Maintenance Operations Request

Request Type	Snow				
Request Date	02/03/2011	Time	3:32:49 PM		
Work Order	3037	Phase	2-3		
Caller Name		Caller Phone		Caller Address	

Street Locations

Street Location	N 67TH ST	Lockey Number(s)	1183,13447,22362
From Intersection	X ST		
To Intersection	BETHANY PARK DR		

Location Address

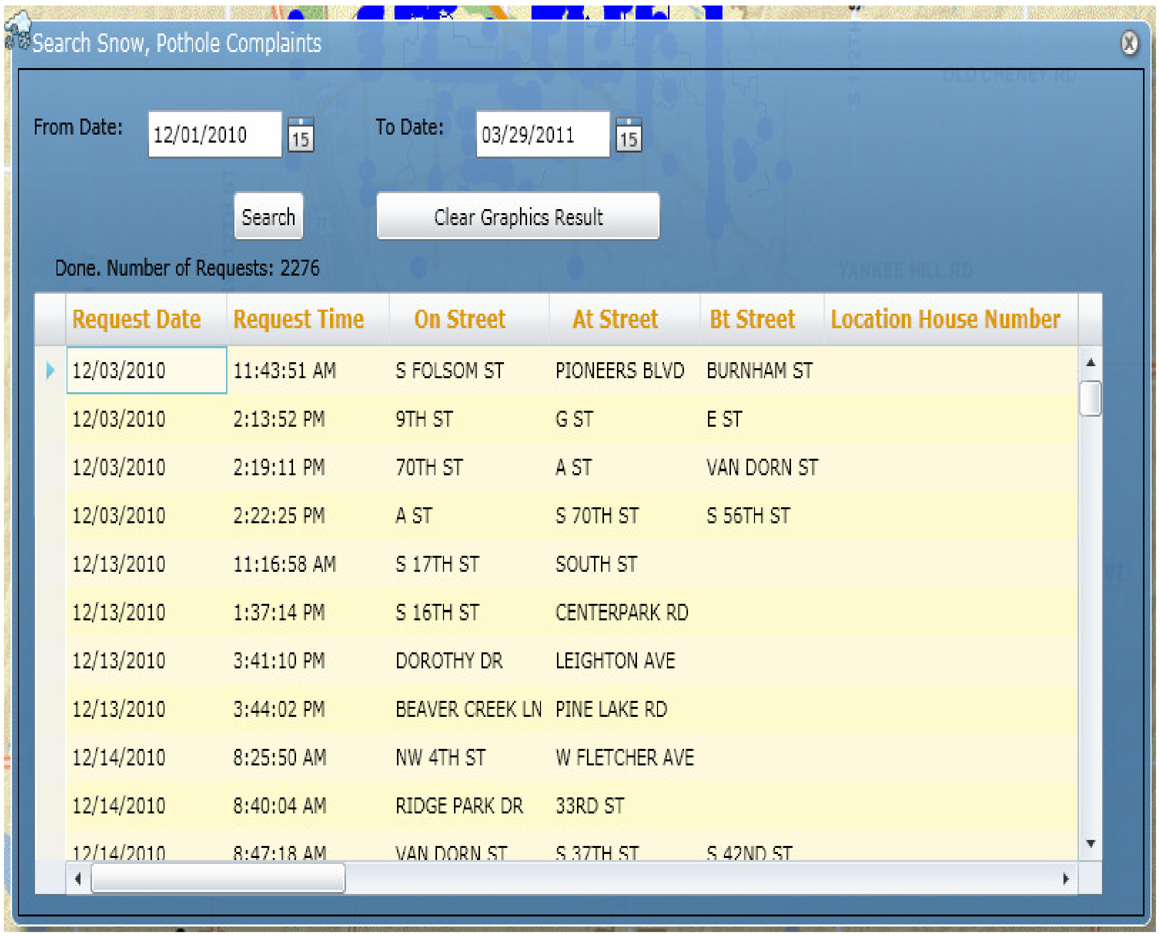
House Number	
Street Name	

Note: Do not enter Street Direction or Suffix

Request Sub Type	Did Not Plow Street
Remarks	
Completion Date	02/04/2011

Figure 9: Snow complaint data entry form.

Second, by geocoding the street location information, we can find the location of a snow complaint and display it in **real-time** on a map as soon as a public works user adds such a complaint to the *StreetManagerDb* database (see Figures 8 and 9). That is a useful new feature because in the past public works database users used a Microsoft Access program to record snow complaints and some days later a GIS specialist may have used street location information to manually geocode the Microsoft Access data using an ESRI ArcGIS geocoding tool. Therefore, the map could not be generated in real-time.



From Date: 12/01/2010 To Date: 03/29/2011

Done. Number of Requests: 2276

Request Date	Request Time	On Street	At Street	Bt Street	Location House Number
12/03/2010	11:43:51 AM	S FOLSOM ST	PIONEERS BLVD	BURNHAM ST	
12/03/2010	2:13:52 PM	9TH ST	G ST	E ST	
12/03/2010	2:19:11 PM	70TH ST	A ST	VAN DORN ST	
12/03/2010	2:22:25 PM	A ST	S 70TH ST	S 56TH ST	
12/13/2010	11:16:58 AM	S 17TH ST	SOUTH ST		
12/13/2010	1:37:14 PM	S 16TH ST	CENTERPARK RD		
12/13/2010	3:41:10 PM	DOROTHY DR	LEIGHTON AVE		
12/13/2010	3:44:02 PM	BEAVER CREEK LN	PINE LAKE RD		
12/14/2010	8:25:50 AM	NW 4TH ST	W FLETCHER AVE		
12/14/2010	8:40:04 AM	RIDGE PARK DR	33RD ST		
12/14/2010	8:47:18 AM	VAN DORN ST	S 37TH ST	S 42ND ST	

Figure 10: Search snow complaints from 12/01/2010 to 03/29/2011.

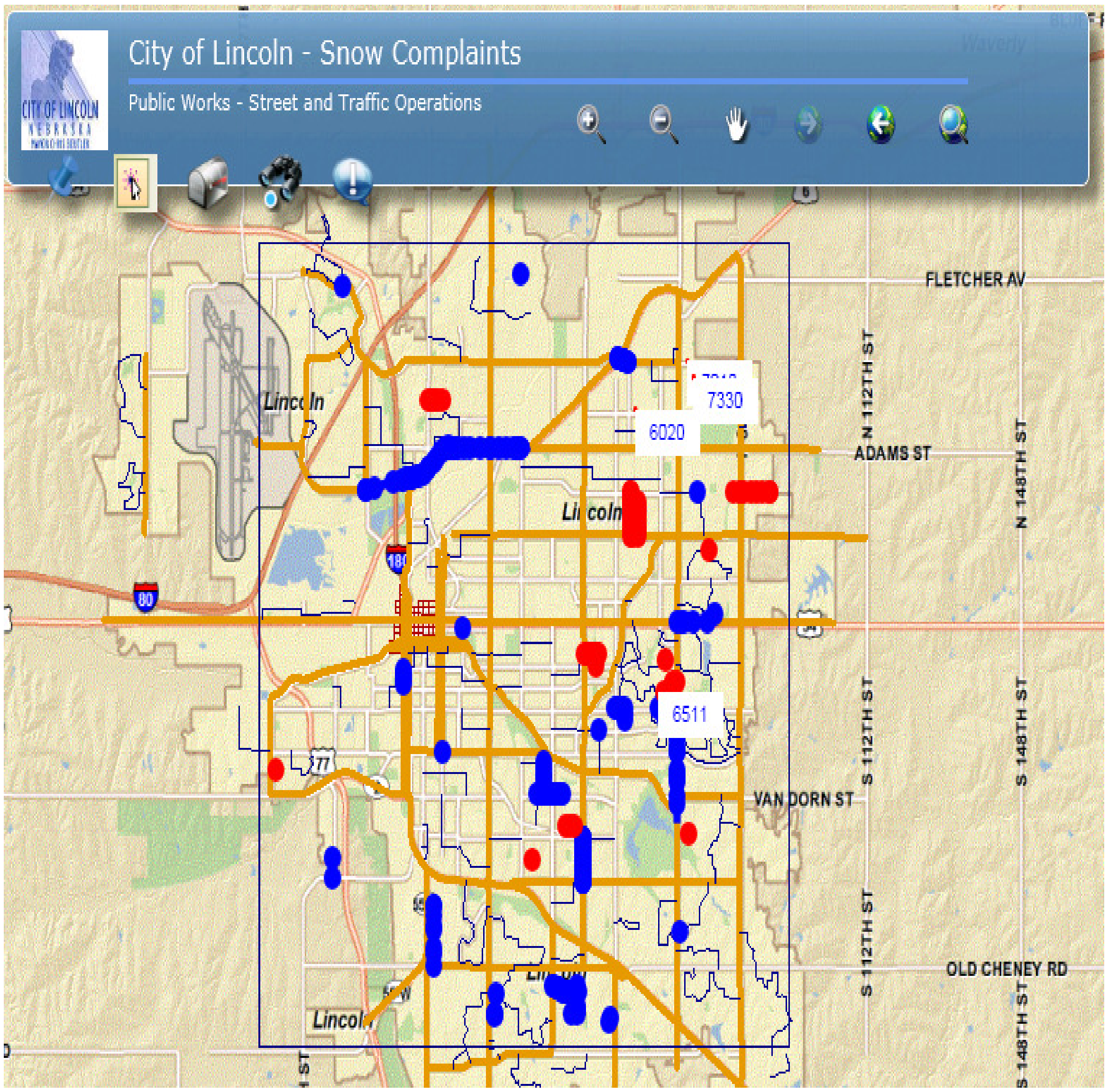


Figure 11: Range queries.

Chapter 6

Efficient Moving Points Estimation

6.1 The Point Estimation Algorithm

As shown in Figure 12, the position of any point P moving linearly along the x-axis can be represented by a function $a_p.t + b_p$ where a_p is the speed and b_p is the starting position of point P. Several GIS applications need to efficiently solve the following:

Count Problem for Moving Points: *Given n moving points P_0, P_1… P_{n-1} in one dimensional space with parametric functions $P_i = a_P.t + b_P$ for $i = 0,…, n-1$ and a query point Q with $Q = a_Q.t + b_Q$, find the number of P_i to the left of Q at given time t.*

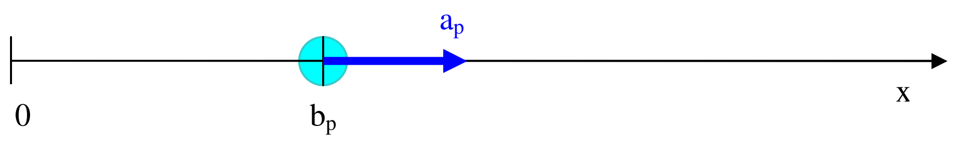


Figure 12: The moving point P starting at b_P and moving with speed a_P.

For example, we may want to know how many cars are to the left of a slow moving vehicle, which may be a snow removal truck.

Revesz [23] presents an indexing algorithm that finds an approximate count for this problem in only O(m log N) time, where m is a constant parameter that controls accuracy of the result and N is the number of moving vehicles. A key element of the algorithm is representing each moving point $P = a_p.t + b_p$ by a static point $P' = (a_p, b_p)$ in a dual plane as shown below in Figure 11.

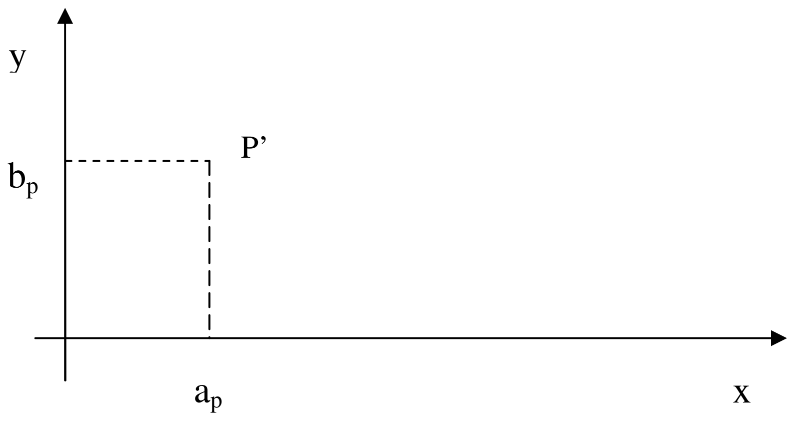


Figure 13: The dual representation of point $P(a_P,b_P)$.

This dual representation is attractive because of the following well-known lemma:

Lemma 1 Let $P = a_p.t + b_p$ and $Q = a_Q.t + b_Q$ be two moving points in one-dimensional space, and let $P' = (a_p, b_p)$ and $Q' = (a_Q, b_Q)$ be their corresponding static points in the dual plane. If P overtakes Q or vice versa at time instance t, then the following must hold:

$$-t = \frac{b_P - b_Q}{a_P - a_Q}$$

That is, -t is the slope of the line P'Q'. Hence, the **Count Problem for Moving Points** reduces to the problem of finding how many points are below a ***query line*** **L**, where **L** is a line crossing Q' with slope –t in the dual plane.

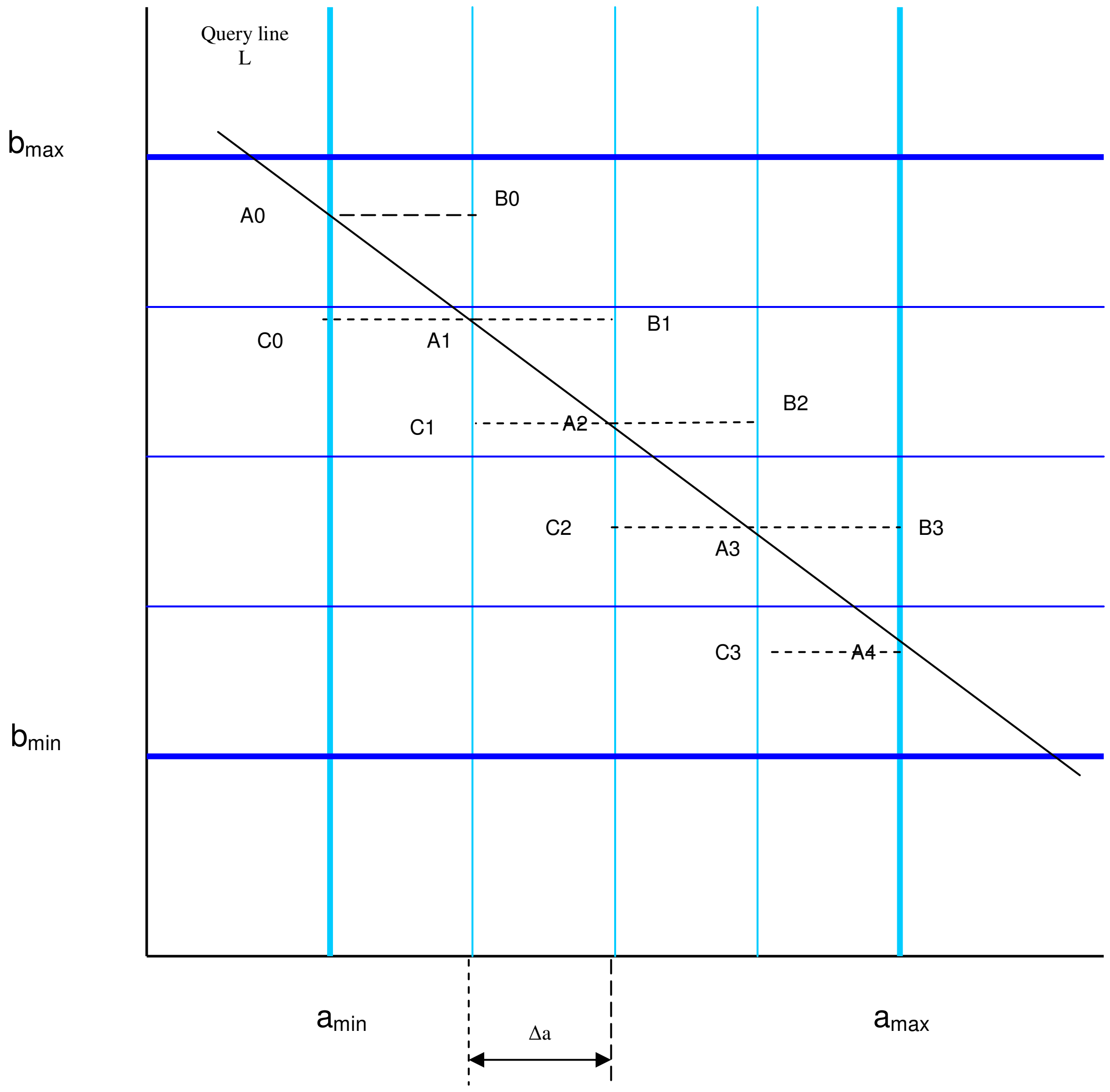


Figure 14: Approximating points below query line L.

Further, [23] reduces the problem of finding the number of points below a line to an O(m) number of ECDF-tree queries as shown in Figure 14, which leads to the following algorithm.

Moving Points Algorithm

Input: -- n points $P_0(a_{P0}, b_{P0}), \ldots, P_{n-1}(aP_{n-1}, bP_{n-1})$

– query point Q (a_Q, b_Q)

– m: the number of line divisions

– t: the query time in seconds

Output: the number of points P_i dominated by Q at time t.

1. Find $a_{max} = \max(a_{p0},\ldots,a_{pn-1})$,

   $a_{min} = \min(a_{p0},\ldots,a_{pn-1})$,

   $b_{max} = \max(b_{p0},\ldots,b_{pn-1})$, and

   $b_{min} = \min(b_{p0},\ldots,b_{pn-1})$.

2. Compute $a = (a_{max} - a_{min})/m$ and

   $b = (b_{max} - b_{min})/m$.

3. Compute arrays A[m], B[m-1], and C[m-1].

4. Compute and return the following approximate count of the number of points below line L through Q with slope –t:

$$Approx\ \#below = \frac{\#(A_1, I) + \#(A_{m+1}, I) + \sum_{i=0}^{m} \#(B_i, I) - \#(C_i, I)}{2}$$

where, I is the ECDF-tree that stores the dual representations of the moving points and #(A_i, I) is the number of points dominated by point A_i, #(B_i, I) is the number of points dominated by point B_i, and #(C_i, I) is the number of points dominated by point C_i.

6.2 Implementation Results

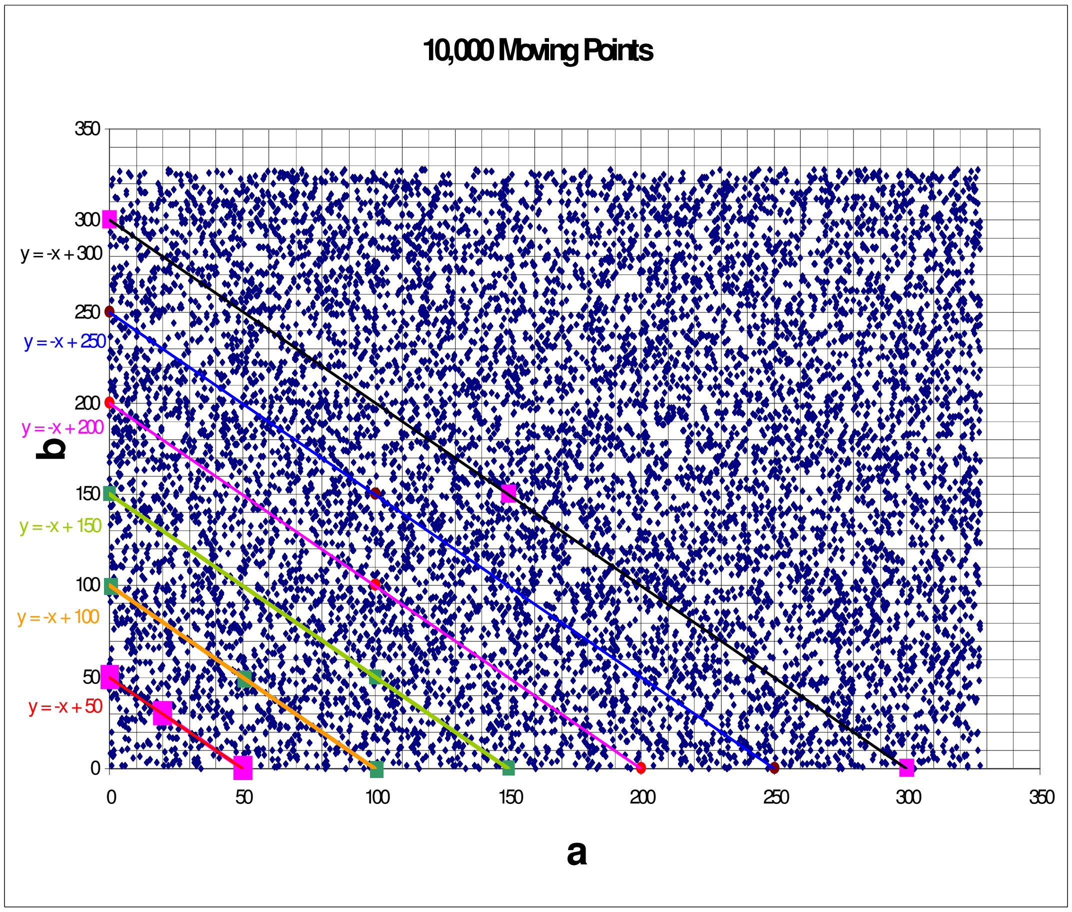


Figure 15: 10,000 moving points and query lines in dual plane.

We implemented the indexing method of [23] on a data set that contained 10,000 random moving points that were created by allowing both a_P and b_P to vary uniformly between 0 and 33. Figure 15 shows a graph of 10,000 moving points and seven query lines in the dual plane.

Table 3 records the running time for counting number of points below each query line L with different m and Q at time t = 1 second. In Table 3 we use the following parameters:

(1) ***# of dominated points***: the number of points dominated by Q at time t.

(2) ***Counting time***: the total of counting time for counting the number points dominated by Q at time t (excluding the time for setting up and creating the ECDF-tree).

The last column of Table 3 is used to show the accurate answers. We can see that the answer with m = 10 had a maximum error of 26, while with m = 100 had a maximum error of only 6. In both cases the maximum error occurred with the query point Q(150,150). For Q(150,150), the error was 0.6 percent with m = 10 and only 0.14 percent with m = 100.

Table 3: Running time results for point dominance queries.

Line		**Estimate with m = 2**	**Estimate with m = 5**	**Estimate with m = 10**	**Estimate with m = 100**	**Estimate with m =10,000**
Q(20,30), t = 1 sec	# of dominated points	356	140	116	103	105
	Counting time	1 msec	1 msec	1 msec	9 msecs	764 msecs
Q(50,50), t = 1 sec	# of dominated points	750	490	445	446	446
	Counting time	< 1 msec	1 msec	1msec	8 msecs	763 msecs
Q(100, 50), t = 1 sec	# of dominated points	1129	1046	1022	1018	1019
	Counting time	< 1 msec	< 1 msec	2 msec	8 msecs	765 msecs
Q(100, 100), t = 1 sec	# of dominated points	2036	1815	1806	1823	1820
	Counting time	< 1 msecs	< 1 msecs	1 msecs	8 msecs	767 msecs
Q(100, 150), t = 1 sec	# of dominated points	3159	2895	2874	2875	2875
	Counting time	1 msec	1 msec	1 msec	8 msecs	770 msecs
Q(150, 150), t = 1 sec	# of dominated points	4324	4174	4176	4144	4150
	Counting time	< 1 msec	< 1msec	1 msec	8 msecs	776 msecs
Q(330, 330), t = 1 sec	# of dominated points	10,000	10,000	10,000	10,000	10,000
	Counting time	1 msec	1 msec	<1 msec	8 msecs	808 msecs

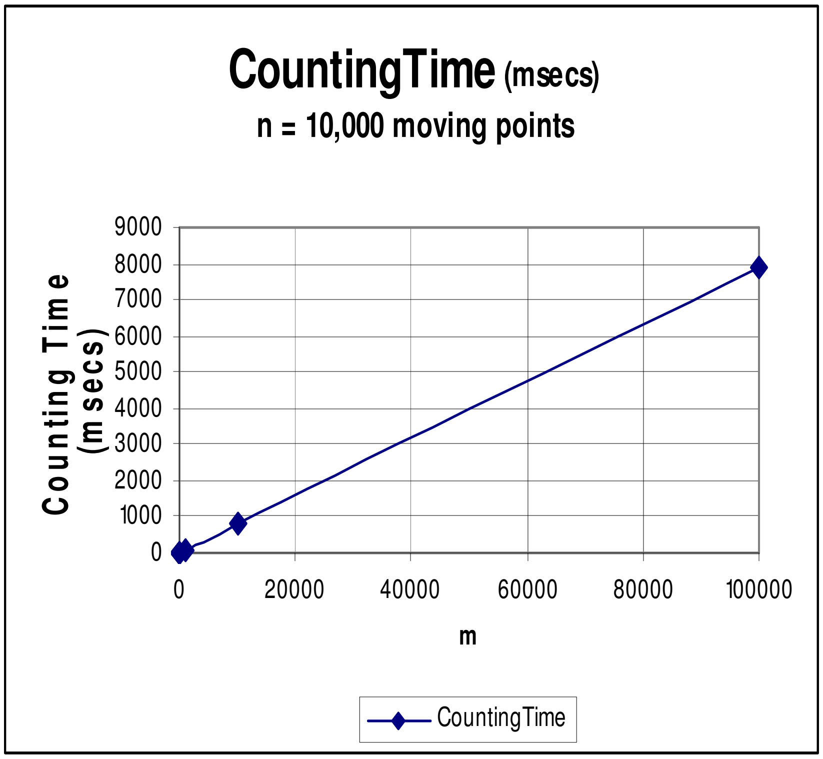


Figure 16: Time vs. m graph with n = 10,0000 moving points.

Figure 16 shows that the counting time varies linearly with the value of m, as expected. Therefore, there is a natural trade-off between the accuracy of the approximation and the counting time.

Chapter 7

Approximated R-Tree

7.1 Virtual point approximation of spatial objects

In some applications, even the bounding object representation of spatial objects is too inefficient. Hence, instead of a bounding box, we developed an approximation of spatial objects by a single point. We call this point a *virtual point* because it is usually not a part of the represented object.

Definition 7.1 Let object P_i be represented by the bounding box with upper-right corner vertex $V_i = (x_i,y_i)$ and lower-right corner vertex $V'_i = (x'_i,y'_i)$. Then V_i is called the *virtual point* of P_i.

Example 7.1 Consider again the street block of Figure 4 that we saw in Chapter 3. From Figure 17, points V_1, V_2, and V_3 are respectively virtual points of land parcels P_1, P_2, and P_3.

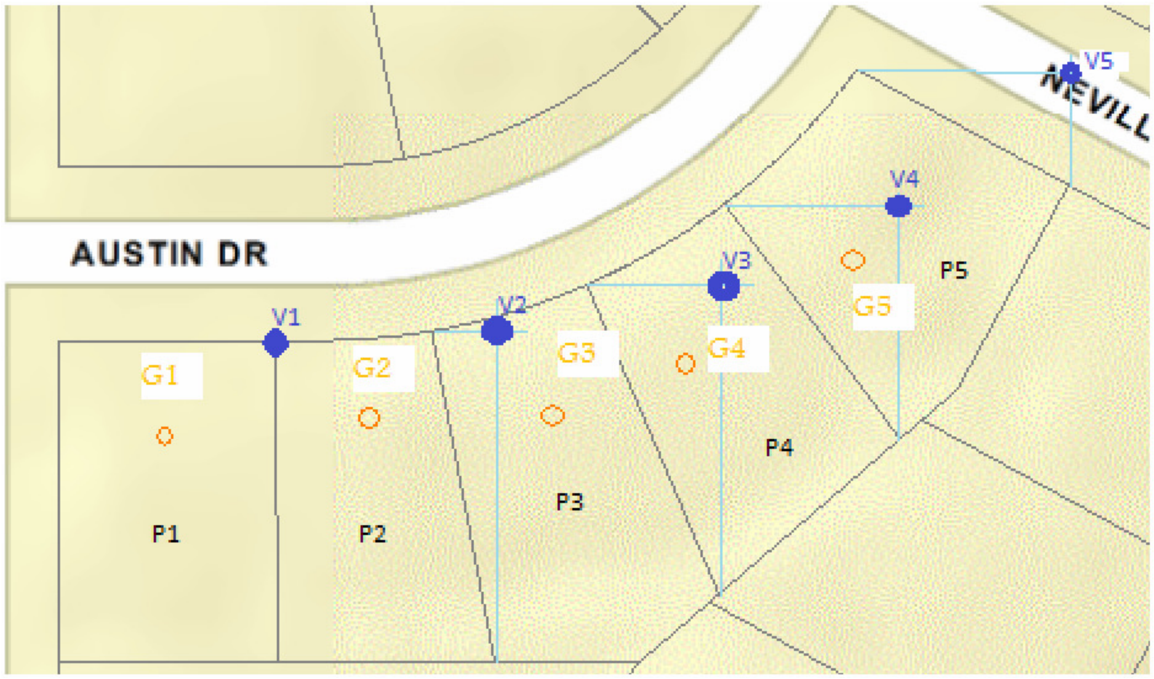


Figure 17: GIS Parcel Polygons' Virtual and GPS Points

The following lemma shows that the dominance property is preserved for virtual points.

Lemma 7.1 Let V_1 and V_2 be virtual points of spatial objects P_1 and P_2, respectively. If V_1 dominates V_2, then V_1 dominates all points of P_2.

Proof: The proof follows from the transitivity of the dominance property.

For example, From Figure 17, polygon parcels P_1, P_2, P_3, P_4, and P_5 are represented by five virtual points V_1, V_2, V_3, V_4, and V_5. Point V_3 dominates the virtual point V_1; thus, V_3 dominates all points of spatial object P_1.

7.2 The Damage Assessment Mobile Application

According to Bolstad [3, page 303], thousands to millions of coordinate pairs or cells are typically required to represent the spatial objects in a GIS. Hence even with modern computers, the retrieval of coordinate data is often too slow. Moreover, many mobile applications running on field devices, such as, Table PCs, Windows CE/Windows Mobile devices, Android phone, or I-phone, have inherent hardware limitations in display, memory, processing power, and battery life.

Consider a Damage Assessment mobile application, where an assessment team needs to record all the collected data into a GIS format and also need to retrieve, view and query damage incidents in real-time. The assessment team may be using a Tablet device with GPS and 3G/4G/WIFI as shown in Figure 18. The Damage Assessment mobile application takes advantage of a GPS signal (using GPS device on a mobile device) to quickly locate a parcel on a GIS map of a mobile device.

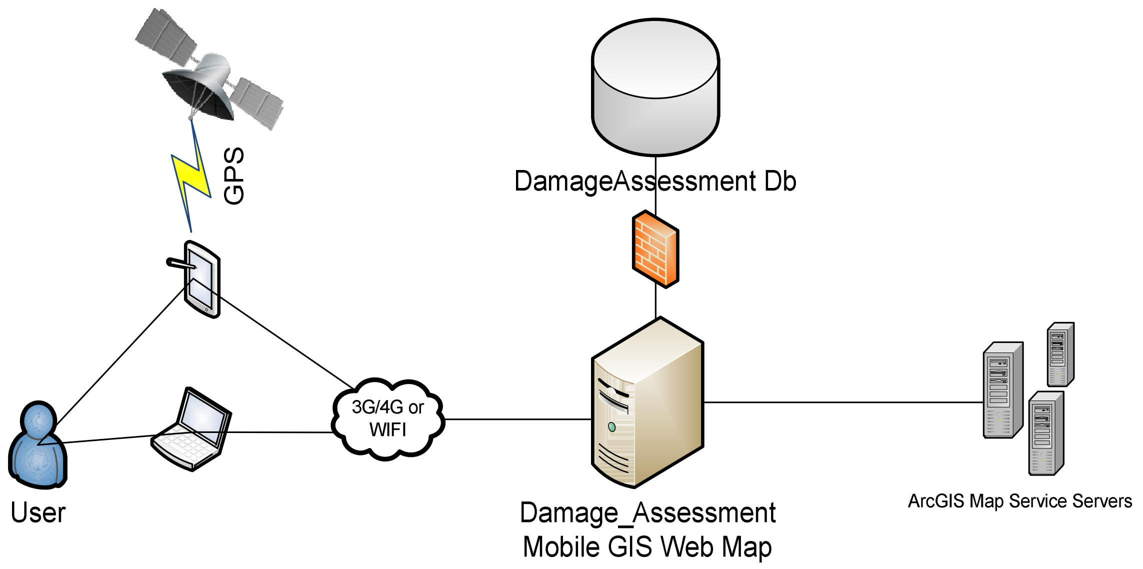

GPS
DamageAssessment Db
3G/4G or WIFI
User
Damage_Assessment Mobile GIS Web Map
ArcGIS Map Service Servers

Figure 18: Damage Assessment GIS Mobile Application Architecture

The Damage Assessment mobile application uses the GIS Parcel layer, which is represented by polygons. Using our approximated method above, we can find all virtual points for representing all parcel polygons. Now the approximated parcel layer contains only points which reduce space significantly and improve performance.

Using virtual points instead of bounding boxes makes the Damage Assessment application easier. A team member standing on a parcel can find a GPS point (x,y), which is part of the parcel. Then we can find the closest virtual point that dominates (x,y). This virtual point represents the parcel, where the field worker is standing. For example from Figure 17, when the field worker is standing on parcel P_1, the GPS point could be G_1, which is dominated by virtual points V_1, V_2, V_3, V_4, V_5, and so on. Out of all those dominating virtual points, V_1 is the closest virtual point. Hence we can estimate that the field worker is standing on parcel P_1. This estimate is not always correct. For example, a worker standing near G_5 may think that he is in parcel P_4 because G_5 is dominated by virtual point V_4, which is closed to G_5. This estimation problem can be reduced if the team worker walks around the parcel and takes several GPS measurements. For example, as the team worker moves to the north and east while staying within parcel P_5, eventually he/she will get a GPS measurement that is no longer dominated by V_4. Then the closest virtual point that always dominates the measured GPS point will be V_5. That would indicate that the team worker is surveying parcel P_5.

7.3 Implementation Results

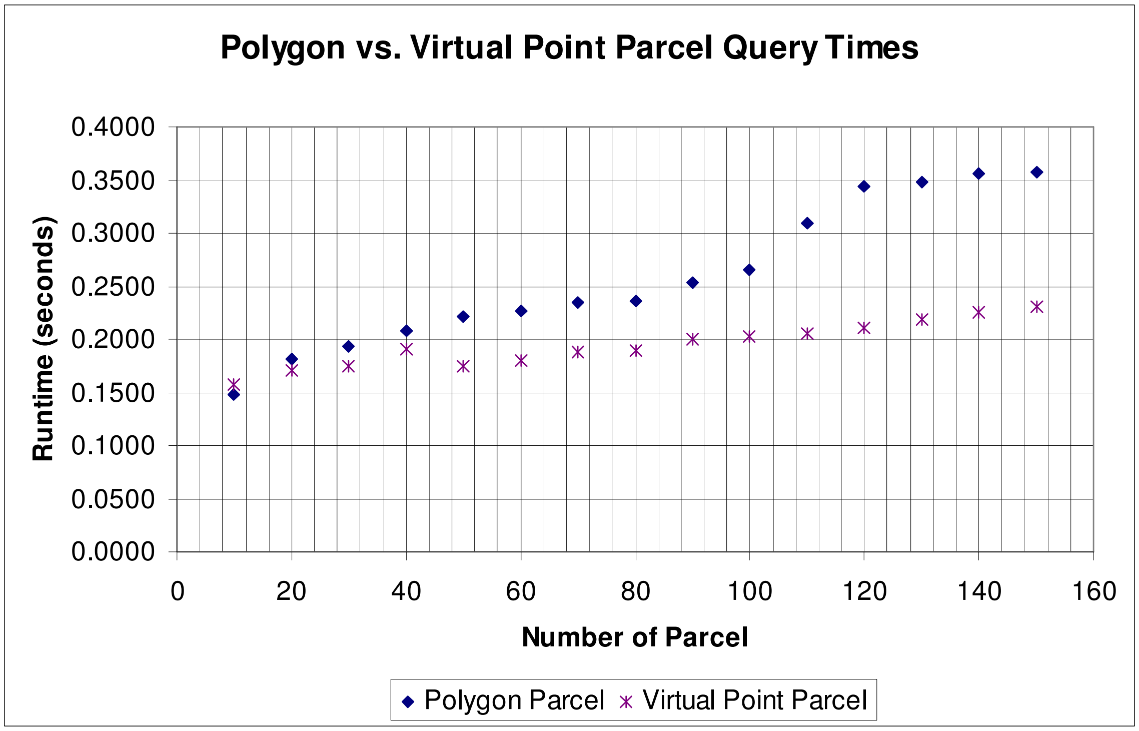


Figure 19: Polygon vs. Virtual Point Parcel Query Times

Using virtual points will improve performance and reduce space with the cost of loosing some information. Figure 19 shows the elapsed time in seconds of a spatial query of polygon parcel versus a spatial query of virtual point parcel.

Chapter 8

Future Work

In the future, we would like to extend our system to show all the moving snow removal trucks. This would be an interesting synthesis of several different types of indexing algorithms. For example, the efficient max-count algorithm of Anderson and Revesz [1] could find the maximum number of snow removal trucks in a rectangular area between two time instances, and the 1-dimensional moving point algorithm could find the number of snow removal complaint locations on a straight road ahead of a particular snow removal truck.

We also plan to consider the problem of estimating aggregate values of the number of moving vehicles in an area when the movement of the vehicles is not linear but polynomial [33].

References

1. Anderson, S. and Revesz, P. Efficient max-count and threshold operators of moving objects. *Geoinformatica*, 13(4): 355-396, 2009.

2. Bao, J., Ying, J., and Hayamizu, S. Development of traffic analysis system using GIS. *Proceedings of the 2nd International Conference on Information Science and Engineering*, pages 36-48, 2010.

3. Bolstad, P. *GIS Fundamentals: A First Text on Geographic Information Systems*, Eider Press, White Bear Lake, Minnesota; 3rd Edition, 2008.

4. Beckmann, N., Kriegel, H., Schneider, R., Seeger, B. The R*-tree: An Efficient and Robust Access Method for Points and Rectangles+. *Proceedings of the 1990 ACM SIGMOD international conference on Management of data,* pages 322-331, 1990.

5. Bentley, J.L.. Multidimensional divide-and-conquer. *Communications of the ACM*, 23(4), 1980.

6. Berchtold, S., Bohm, C., Keim, D.A., and Kriegel, H.P. A cost model for nearest neighbor search in high-dimensional data space. *In PODS*, pages 78-86, 1997.

7. Bigham, J.M., Rice, T.M., Pande, S., Lee, J., Park, S.H., Gutierrez, N., and Ragland, D.R. Geocoding police collision report data from California: A comprehensive approach. *International Journal of Health Geographics*, Volume 8, 2009.

8. Brinkhoff, T., Horn, H., Kriegel, H.P., Schneider, R. A Storage and Access Architecture for Efficient Query Processing in Spatial Database Systems. *Proceedings of the Third International Symposium on Advances in Spatial Databases*, pages 357-376, June 23-25, 1993.

9. City of Lincoln, Criminal justice information system (CJIS): http://cjis.lincoln.ne.gov/~lpd/cfstoday.htm

10. City of Lincoln, 2008 crash study: http://lincoln.ne.gov/city/pworks/engine/crash/index.htm

11. Fang, Y., Friedman, M., Nair, G., Nair, G. Spatial Indexing in Microsoft SQL Server 2008. *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1207-1216, 2008.

12. Guttman, A. R-Trees: A Dynamic Index Structure for Spatial Searching. *Proceedings of the ACM SIGMOD Int. Conf. on management of data*, pages 47-57, 1984.

13. Hjaltason, G.R., Samet, H. Ranking in Spatial Databases. *Proceedings of the 4th International Symposium on Advances in Spatial Databases*. Pages 83-95, 1995.

14. Kamal, M.A. City planning and development using geographic information systems. *Proceedings of the 11th International Conference on Computer and Information Technology*, page 3, 2008.

15. Kanellakis, P., Kuper, G., and Revesz, P., Constraint query languages, *Journal of Computer and System Sciences*, 51(1):26-52, 1995.

16. Kanth, R., Agrawal, D., and Singh A. Dimentionality Reduction for Similarity Searching in Dynamic Databases. *SIGMOD Conference* 1998: 166-176.

17. Kanth K., Ravada, S., Abugov, D. Quadree and R-tree Indexes in Oracle Spatial: A Comaprison using GIS Data. *Proceedings of the ACM SIGMOD Conference*, pages 546-557, 2002.

18. Kriegel, H.P., Schiwietz, M., Schneider, R., Seeger, B. Performance comparison of point and spatial access methods. *Proceedings of the 1st Symposium on Design and Iimplementation of Large Spatial Databases*. Springer Lecture Notes in Computer Science, pages 89 – 114, 1989.

19. Ngo, A. Revesz P. Efficient Traffic Crash and Snow Complaint GIS System, *Proceedings of the 12th International Conference on Digital Government Research*, ACM Press, 10 pages, College Park, MD, USA, 2011.

20. Nwaneri, S.O. Mapping intersection accidents with GIS technology in Huntsville, Alabama, U.S.A. *Proceedings of the IEEE International Geoscience and Remote Sensing Symposium*, volume 6, pages 3727-3729, 2003.

21. Papadias, D., Theodorisdis, Y., Sellis, T., and Egenhofer, M. Topological relations in the world of minimum bounding rectangles: a study with R-trees. *Proceedings of the ACM SIGMOD Conference*, San Jose, California, 1995.

22. Parrish, L.S., Dixon, B., Cordes, D., Vrbsky, S.,and Brown,.D. CARE: An automobile crash data analysis tool. *IEEE Computer Society*, 36(6):22, June 2003.

23. Revesz, P. Efficient rectangle indexing algorithms based on point dominance. *Proceedings of the 12th International Symposium on Temporal Representation and Reasoning, IEEE Press*, pages 210-12, Burlington, VT, June 2005.

24. Revesz, P. *Introduction to Databases: From Biological to Spatio-Temporal*, Springer, 2010.

25. Samet, H. *Foundations of Multidimensional and Metric Data Structures*, Morgan-Kaufmann, San Francisco, 2006.

26. Samet, H. *The Design and Analysis of Spatial Data Structures,* Addison-Wesley, Reading, MA, 1990.

27. Samet, H. *Applications of Spatial Data Structures: Computer Graphics, Image Processing, and GIS*. Addison-Wesley, Reading, MA, 1990.

28. Seeger, B., Kriegel, H. Design and implementation of the buddy tree. *Computer Science Technical Report* 3/90, University of Bremen, submitted for publication, 1990.

29. Sellis, T., Roussopoulos, N., Faloutsos, C. The R+-Tree: A Dynamic Index for Multi-Dimensional Objects. *Proceedings of 13th International Conference on Very Large Data Bases*, pages 507-518, 1987.

30. Sewell, C. Remembering a snowstorm that paralyzed the city, *New York Times*, February 10, 2009.

31. Steenberghen, T., Dufays, T., Thomas, I., and Flahaut, B. Intra-urban location and clustering of road accidents using GIS: A Belgian example. *International Journal of Geographical Information Science*, 18(2):169–181, 2004.

32. Steiner, R., Bejleri, I., Yang, X., and Kim, D. Improving geocoding of traffic crashes using a custom ArcGIS address matching application. *Proceedings of the 22nd Environmental Systems Research Institute International User Conference*, San Diego, 2003.

33. Yue, H., Jones, E., Revesz, P., Local polynomial regression models for vehicle speed estimation and forecasting in linear constraint databases. *Proceedings of the 17th International Symposium on Temporal Representation and Reasoning*, IEEE Press, pages 154-161, Paris, France, September 2010.

34. Yuval, G. Finding Near Neighbors in k-dimensional Space. *ACM Transactions on Mathematical Software*, Volume 5, Issue 2, pages 183-192, 1979.

35. Zhang, D., Tsotras, V. J., Gunopulos, D., Efficient aggregation over objects with extent, *Proceedings of the 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 121-132, 2002.

36. Zhang, J., Chen, L., You, S., and Chen, C. A hybrid approach to segment-type geocoding of New York City traffic data. *Proceedings of the 1st International*

Conference on Computing for Geospatial Research and Applications, ACM Digital Library, http://dx.doi.org/10.1145/1823854.1823871, 2010.