# A Hybrid Approach to Segment-Type Geocoding of New York City Traffic Data

Jianting Zhang
Department of Computer Science
City College of New York
New York, NY 10031
jzhang@cs.ccny.cuny.edu

Simin You
Department of Computer Science
City University of New York
Graduate Center
New York, NY, 10006
syou@gc.cuny.edu

Li Chen
Department of Civil Engineering
City College of New York
New York, NY 10031
lchen@gc.cuny.edu

Cynthia Chen
Department of Civil and Environmental Engineering
University of Washington
Seattle, WA 98195
qzchen@u.washington.edu

## ABSTRACT

Many types of traffic data are often recorded as (Main street, From street, To street) triples. All the segments between the intersection pairs of (Main street,From street) and (Main street,To street) need to be geocoded, with additional directional constraints. We term this new type of geocoding task as segment-type gecoding in contrast with classic geocoding that takes a street address or intersection and converts the address to a pair of coordinates. Most of the existing geocoding software does not have the capability to handle such segment-type geocoding. This motivates us to develop algorithms and programs for the new type of gecoding in the context of the Effectiveness of Traffic Calming study sponsored by the New York City Department of Transportation (NYCDOT). Due to the level of complexity of the New York City street network, we have adopted a hybrid approach. The hybrid approach includes several algorithms to automatically geocode well-formed traffic records and a software tool built on top of ESRI ArcMap to facilitate manual geocoding of ill-formed traffic records. The hybrid approach has achieved desired accuracies with reasonable manual involvements. We believe the approach is applicable to similar projects that involve segment-type geocoding tasks.

## Keywords

Geocoding,Street Segments, Traffic Data, NYC

## 1. INTRODUCTION

Geocoding can have multiple meanings under different contexts. According to Wikipedia [5], "Geocoding is the process of finding associated geographic coordinates (often expressed as latitude and longitude) from other geographic data, such as street addresses, or zip codes (postal codes)." We refer readers to [10] for surveys and more comprehensive discussions on different aspects of geocoding. Gecoding serves as one of the major data collection approaches to Geographical Information System (GIS) where geocoded data can be further analyzed. Geocoding techniques have been widely applied in many fields including health [8][15][11] and transportation [16][17]. While major commercial GIS software such as ArcGIS [1] and MapInfo [4], have provided geocoding functionality for decades, Google Map and other Internet-based mapping software have made the technologies virtually freely available to anyone. Subsequently, geocoding has gained increasing popularity over the past few years. This in turn has spawned more applications, especially Web-based applications.

While many applications use geocoding to convert text-based address data to geographical coordinates, in this study, we address a non-conventional geocoding task that requires finding segments along a polyline between two nodes in a network. The task arises from handling traffic records in our Traffic Calming Effectiveness project where treatment, speed and volume records are given in the form of (Main street, From street, To street) triples and we need to find all the street segments along the Main street that are between the From street and the To street. As the coordinates of the street segments are already known in a given a street network, the street segment identifiers are sufficient in most of subsequent analyses and there is no need to further convert the segments to a collection of coordinates. To distinguish this new type of geocoding task with existing ones, we term the new types of geocoding as the segment-type geocoding. It is not difficult to see that the segment-type geocoding is an extension of traditional intersection based node-type geocoding. Unfortunately, to the best of our knowledge, the new type of geocoding task is not supported by any known commercial or open source GIS or gecoding software, which motivates us to develop our own techniques for the project.

Although the task arises from the domain of handling certain types of road traffic data, we believe potentially the proposed technique can be applied to a variety of other domains. For example, geocoding bus/railway travel paths in the form of (route#, source, destination) triples. Another example would be geocoding travel survey data in the same form of (Main street, From street, To street) triples. With a robust and efficient segment-type geocoder, for geographi-



*COM.Geo 2010*, June 21-23, 2010 Washington, DC, USA

cal data derived from a common network infrastructure, it is possible to compress long polylines that are normally represented as a collection of points with a triple of (Main street ID, From street ID, To street ID). This may significantly save data storage and make the related applications more portable.

Being one of the largest cities in the world, the New York City (NYC) has a complex street network. The 09C release of NYC Department of City Planning (DCP) LION dataset [2] has 11,705 standardized unique street names and 175, 440 segments. We note that many street names in NYC have alternative names due to historical,political and cultural reasons making geocoding NYC traffic data more difficult. Finally, some of the traffic data come from New York State Department of Transportation (NYSDOT) using state highway system rather than NYC street network. Geocoding the NYSDOT data to NYC LION segments is very challenging since the two infrastructure systems are involved. More importantly, issues related to scale mismatch may arise. Given the problem complexity and time constraints, we have decided to adopt a hybrid approach that integrates algorithm-based automatic approach with software tool-assisted manual approach. Results have shown that the hybrid approach is very successful in geocoding nearly thirteen thousand traffic records within time and budget constraints. Our technical contributions can be summarized in the following:

- First, we have developed a set of algorithms to automatically geocode traffic records in the form of (Main street, From street, To street) triples. Unpromising records are automatically marked for manual geocoding.
- Second, we have developed a Visual Basic for Application (VBA) based ArcGIS tool to facilitate manual geocoding by providing customized graphical user interfaces and specialized functions. Software-assisted manual geocoding has significantly increased efficiency.
- Third, we have applied the hybrid approach to different sources of traffic data, including treatment, volume and speed data. Manual validations based on a rigidly designed statistic framework have shown satisfactory results.

The rest of the paper is arranged as the following. Section 2 introduces the background and related work. Section 3 presents the hybrid approach to the segment-type geocoding task. Section 4 describes the procedures to evaluate the accuracies of the hybrid approach and demonstrates its effectiveness. Finally Section 5 is the conclusion and future work discussions.

## 2. BACKGROUND AND RELATED WORK

Geocoding is the act of transforming aspatial locationally descriptive text into a valid spatial representation using a predefined process [11]. While Global Positioning System (GPS) devices that can measure latitude and longitude coordinates are becoming increasingly available and affordable, many geo-referenced data are still being generated in textual format. Many achieved geo-referenced data, including traffic records, are often available in the form of descriptive text. In order to import these data into GIS and/or other information systems, it is necessary to transform the descriptive text into coordinates or their alike so that sophisticated analysis, such as spatial statistics and spatial clustering based regionalization, can be performed beyond full text search of the textual descriptions.

Existing literature shows that health related research [8][15][11] and social-economic studies [16][17] heavily rely on gecoding techniques where postal addresses of study objects need to be converted to geographical coordinates. Geocoding techniques have also been applied to study road accidents [20] [19] [7]. However, police reports on road accidents are seldom provided as postal addresses; rather, they are more likely to be provided as street intersections (with or without an offset) on local streets or highway postmile marker (or alike) on highways. Most existing studies rely on commercial GIS software to provide geocoding functionality. Indeed, for geocoding street intersections, existing GIS software works excellently for postal addresses provided that an accurate base map is available and the linear interpolation assumption is satisfied. However, in complex street networks, an intersection of two streets may result in multiple nodes. GIS software, e.g., ArcGIS, often picks one while treating the rest as ties which is unnatural [7]. While we do not address issues related to the node-type geocoding for NYC in this paper, we note that there could be as many as 25 nodes corresponding to an intersection in NYC. This is not uncommon if both streets have one central line, two roadbeds and two service roads. In addition, when commercial GIS software does not support geocoding intersections with offset, users will have to develop their own modules on top of GIS geocoding modules [20][7]. Another drawback of commercial GIS geocoding software is that they usually do not support correcting geocoding errors interactively through manual processes as reported in [20][7].

The work reported in [7] is dedicated to geocode police collision report data in California. For a total of 142,007 fatal and severe injury collisions identified in California Statewide Integrated Traffic Records System (SWITRS) from 1997 to 2006, they have achieved 99.8% for postmile-coded collisions and 86% for intersection-coded collisions. Among the intersection-coded collisions, 65% of them occurred at some distance from the intersection, i.e., with an offset. For postmile collisions and intersection collisions with offset, the authors have built customer modules to integrate linear referencing with gecoding in ArcGIS as the function is not directly supported by the software. Our work on developing tool-assisted manual geocoding program on top of ArcGIS follows a similar strategy. We have not utilized geocoding functionality in ArcGIS as it does not support segment-type geocoding. However, we do take advantages of graphics related operations (zoom and pan) and table query (street name matching) provided by ArcGIS to build our tool.

In addition to [10], quite a few studies on evaluating geocoding accuracies and/or uncertainties have been reported [13][12][24][18][23]. Studies show that geocoding accuracies may vary due to the use of different algorithms implemented in commercial software and various degrees of accuracies of underlying transportation networks. As most software assume the coordinates of the address can be linearly interpolated between the coordinates of two known addresses that the address to be geocoded fall in between, the accuracy also relies on how well the assumption can be satisfied. The work reported in [23] compares the classic street network based geocoding with two alternative approaches, i.e., gecoding using a master address file and geocoding using land parcel data. His results suggest that using a master address file can improve the geocoding accuracy while using land parcel data may decrease accuracy. These studies have illustrated the accuracies and/or uncertainties associated with geocoding. However, no previous studies have addressed issues on how to evaluate accuracies for segment-type geocoding.

We want to mention that studies related to geocoding are also proliferating in quite a few computer science fields, including spatial databases and information retrieval. While traditionally spatial databases and GIS are designed to manage numeric coordinates, the work on spatio-textual spreadsheet reported in [14] specifies spatial attribute values textually. In addition, string and spatial coordinates have been co-indexed for fast approximate string search [22]

which has great potential to improve geocoding accuracy. A large portion of the papers appeared in the proceedings of the American Computing Machinery (ACM) workshop on Geographic Information Retrieval (GIR) have explicitly addressed gecoding issues, e.g., [6][21]. However, similar to geocoding in health and transportation applications, these studies are limited to geocoding postal addresses. We hope our work on segment-type geocoding can contribute to spatial database and information retrieval research.
Finally we notice that Geosupport Desktop Edition is a highly customized geocoding package that allows users to process geographic information for New York City[3]. Among the variety of functions it provides, function "3S" is the one that similar to the segment type-geocoding most. While Geosupport is able to find stretches in simple cases, very often it fails to automatically geocode streets that have multiple roadbeds. For example, Geosupport can not process record (Queens Blvd, Union TP, Hillside Av) and reports an error on Queens Blvd and Union TP intersect more than twice. Since Geosupport is not open source and requires a license, we have decided to develop our own algorithms to solve the problem.

# 3. METHODOLOGY

There are three types of traffic datasets that require segment-type geocoding in our project, i.e., traffic claming treatment, traffic volume and speed. A large portion of NYC street segments have multiple roadbeds and some major roads also have service roads (with the same street names) in both directions, in addition to an imaginary central line. As shown in Fig. 1, there could be up to five parallel street segments for a (Main street, From street, To street) triple, even if the intersections of (Main street, From street) and (Main street, To street) are neighboring intersections along the Main street. Finding all the relevant segments (which are required by modeling traffic calming measurements) is significantly more difficult than node-type geocoding. Some traffic datasets also have a direction requirement which is given as the overall direction of major roads rather than individual street segments. The overall direction restriction, which may be different from the individual segment direction, further complicates the segment-type geocoding tasks. In addition, some streets are interrupted by some private building blocks such as large hospitals and college campuses. The disconnected network topologies have made several algorithms that we tried during the early stage of development not working properly.

Our approach to geocoding NYC segment-type traffic data is a combination of algorithm-based automatic approach and software tool-assisted manual approach. In this section, we will first introduce the data model of the LION dataset which serves as the underlying infrastructure for our geocoding. Our data cleaning (or preprocessing) stage also heavily relies on the standardized street names included in the LION dataset and its associated alternative street name table. We then present our algorithms to automatically geocode segment-type traffic data. The VBA tool to facilitate manual geocoding for failed (or ill-formed) records in the automatic geocoding step is subsequently presented.

## 3.1 The LION Dataset and Data Cleaning

The first step for segment-type geocoding is to map the street names given in a (Main street, From street, To street) triple to street identifies based on which the segments of the street can be retrieved and subsequently filtered and ordered (see next subsection for more details). By looking into the traffic datasets we have found the following problems in mapping street names to their identifies. First of all, street names are often misspelled. The possible ways of

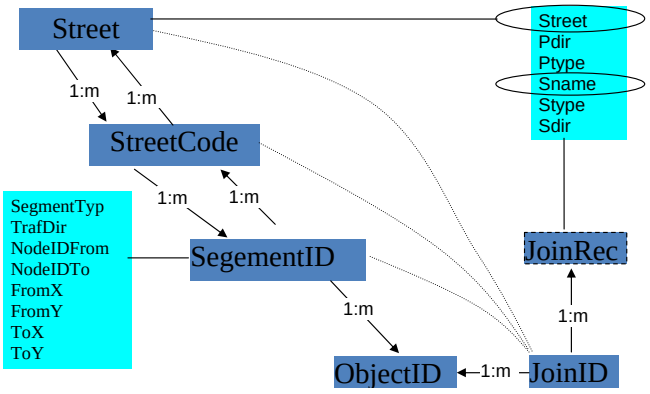


**Figure 2: LION Data Model Relevant to Segment-Type Geocoding**

misspelling (including using non-standard abbreviations) are only limited by imaginations. Second, quite some wrong information are provided in the street names, such as direction, borough and suffix. In the Borough of Queens, numbered street names may end with "ST", "AV","RD" and "PL" and it is not uncommon that they are incorrectly specified. A common problem in the Borough of Manhattan is that directions of streets are missing while two streets may differ only by directions. Third, some streets may have multiple names for certain portions due to historical, political and cultural regions. To solve the problems, we match the street names in traffic records with the standardized names in the LION dataset to be detailed below. As the LION dataset is the base map (or infrastructure) for our geocoding, part of the data model that is relevant to our geocoding task is provided in Fig. 2.

From Fig. 2 we can see that, different from our expectation, the relation between street name (Street) and its identifier (StreetCode) is n:m rather than 1:1 which means that a street name can have multiple street codes and a street code may correspond to multiple street names. Quite a few segments (with same SegmentID values) have been duplicated in the LION dataset to allow a same segment to associate with multiple JoinID values through the unique ObjectID. The data model allows a street segment to have multiple alternative names given in the altname table. While a SegmentID value may have multiple ObjectID values, the geometric and transportation related features are unique which allows use SegmentID as the basic unit. A join record in the altname table has six components, namely PDir (prefix direction), PType (prefix type), SName (street name-main body), SType (suffix type), Sdir (suffix direction) and Street (combined). Note that the Street field has all the legal alternative street names which makes its values a superset of the Street field values in the LION dataset. According to the LION documentation, while the altname table was originally designed to facilitate classic node-type geocoding in ArcGIS, we have learnt from the design and used the following approach to match the street names in the traffic records with the street names in the Street field of the LION dataset. First, we parse the street names in both data sources into five components as in a join record in the altname table, i.e., PDir, PType, SName, SType and Sdir. We then use SName as the primary component and the rests as the secondary components to compute matching scores. Finally, for the pair with the largest score, if the score is above a certain threshhold, a matching decision will be made; otherwise, the street name in the traffic record will be left

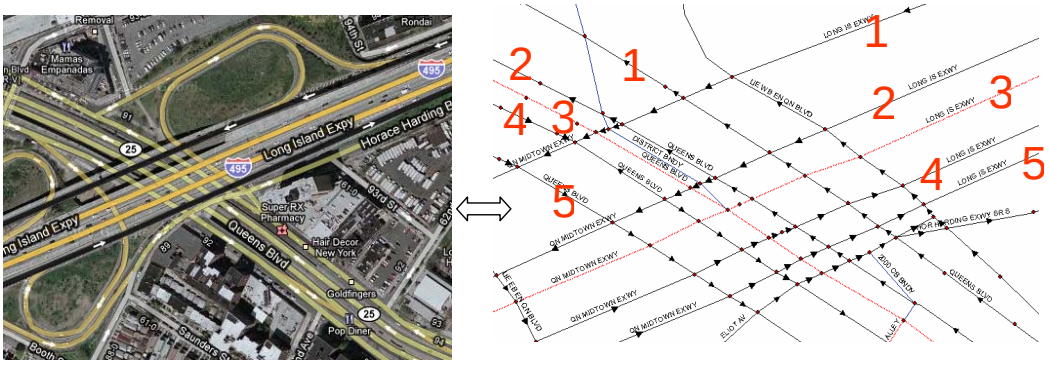


**Figure 1: An Example illustrating the Complexity of NYC Street Network**

for manual checking and corrections. For a traffic record whose three street names in its triple are all matched with the Street field in the LION dataset with or without manual correction, i.e., well-formed records, they will be sent for automatic gecoding as detailed in Section 3.2. The rest records, i.e., ill-formed records. will either be marked as non-geocodable or sent for manual geocoding where street names can further be corrected using a map-based visual inspection approach as detailed in Section 3.3.

## 3.2 Algorithm-based Automatic Geocoding

Given a (Main street, From street, To street) triple, the algorithm finds all the street segments between the From street and To street and along a Main street. For notation convenience, the three streets are abbreviated as MainST, FromST and ToST, respectively. We also term a collection of connected segments as a street stretch. The algorithm has two components: the first is geometric filtering to find candidate segments and the second is to remove false positives through chaining connected segments.

**Geometric Filtering.** The geometric filtering algorithm together with an example are provided in Fig. 3.The bounding box of the nodes in the (MainST, FromST) and (MainST, ToST) intersections are first calculated. Only the segments in the MainST that fall within the bounding box are considered for further processing in the next step. The filtering algorithm may cause false positives and true absences as indicated by the red circles and curves in Fig. 3. However, we argue that the chance of true absences is very slim as the streets like the dotted curves in Fig. 3 are very rare. We can use a larger expansion rate of the bounding box (Step 3) to further reduce the probability. In the worst case, we can skip the geometric filtering step and use all the segments associated with the MainST (Step 4) as the candidates for further processing as described next.

**Chaining Connected Segments.** The algorithm to chain connected segments or street stretches is outlined in Fig. 4. Step 3 plays a key role in the algorithm. For simple intersections that have only one node (Fig. 5A), this step actually can be skipped because it takes no effect. However, for intersections that have multiple nodes (Fig. 5B and Fig. 5C), since it is difficult to provide a total order of these nodes (based on either geometry or topology) which is required, the nodes in the middle of the (MainST, FromST) intersection may be the first ones to be used as the starting nodes in step 6.1. This is problematic as tracing on both directions are allowed by the algorithm. As a consequence, the false positives segments are included in the segment chaining process which is undesirable (the dotted red lines in Fig. 5B). On the other hand, if we separate the segments in the two intersections from the rest (as shown in Fig. 5C), although all the nodes are en-queued in Step 6.5, the intersection nodes can be skipped during the chaining process (Step 6.1) as they are not connected to any segments in the reduced_ids. The segments that the right-most nodes are connected (highlighted in solid light gray lines) follow the desired direction and will be used as the seed segments for further chaining. In contrast, the segments that are connected by the left-most nodes (highlighted in dotted red lines) do not follow the desired direction. They are correctly treated as false positives and subsequently removed. The segments within the two intersections can be added back to the output if needed (Step 8). Note that we have skipped Step 7 due to space limit.

Another important issue is how to determine the appropriate direction to follow druing the chaining process. In real street networks, especially the network of NYC, roadbeds may split, merge and change directions. These might be problems in chaining segments into stretches. Our solution is to apply a combination of absolute direction, relative direction and node degree constraints. The absolute direction of a street stretch can be computed based on the average coordinates of nodes in the (MainST, FromST) intersection and (MainST, ToST) intersection, respectively. If the angle is too big (currently the threshold is set to $\pi/2$), then the segment can be a potential false positive (Fig. 6A). However, unless a stretch has only one segment, the angle may not represent the actual street stretch direction. When multiple segments are chained, we compute the relative direction between two consecutive segments instead (Fig. 6B). When the algorithm detects that there is a sharp direction change (currently the threshold is set to $\pi/3$), the chaining process will stop at the node connecting the two segments, if the second segment is the only one connecting the first segment. On the other hand, when there are multiple segments connect to a previous segment through the middle node, then the segment that has the smallest angle with the previous segment AND the angle is less than a threshold (currently set to $\pi/10$) will be chosen and the middle node will be added to the queue. This will allow the algorithm start at the middle node to chain additional segments into stretches (Fig. 6C). Otherwise the middle node will be considered as a stopping node and will be added to the queue without further chaining (Fig. 6D). If the thresholds are set larger, more traffic records will pass through the automatic geocoding process and fewer records will be sent for manual geocoding, however, at the risk of less accuracy.

## 3.3 VBA Tool to Facilitate Manual Geocoding

1 Retrieve nodes in the MainST & FromST and MainST&ToST intersections.
2 Find the bounding box of the combination of the nodes in the two intersections.
3 Expand the bounding box as necessary
4 Retrieve all the segments associated with the MainST.
5 Filter the segments using the bounding box.

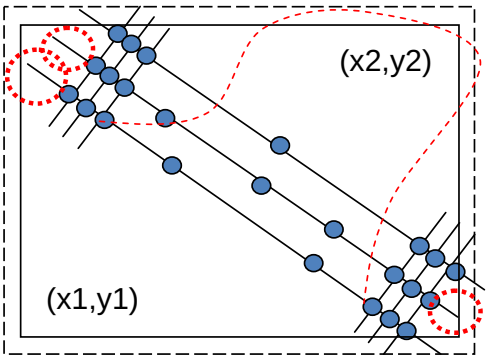


**Figure 3: Algorithm for Geometric Filtering**

Input: MainST, FromST, ToST, Boro, Set of Segment IDs resulting from the geometric filtering component (geo_ids)
Output: Set of Segment IDs between FromST and ToST in Boro and satisfying optional direction constraints.
1 Retrieve the from/to intersection node sets.
2 Find all segments that connect nodes in two intersections, respectively
2.1 For each of the nodes in the FromST (or ToST) node set
2.1.1 Find all the segments associated with the node.
2.1.2 Only keep the segments that are in the segment set of the MainST.
3 Remove the segments in from/to intersections from geo_ids and call it reduced_ids.
4 Initialize a queue data structure with the nodes in the MainST/FromST intersection.
5 Initialize a Set data structure (temp_ids) with all the segments in reduced_ids.
6 While the queue is NOT empty
6.1 De-queue a node and use it as the starting node.
6.2 Use a Depth-First Search (DFS) to chain connected segments and grow a stretch.
6.3 Remove the chained segments from temp_ids.
6.4 The segment chaining process stops at certain nodes based on the stopping criteria which may be a combination of direction and topology.
6.5 The stopping node is added to the queue data structure.
7 For each of the chained stretches
7.1 If there are direction constraints (e.g., E/W bounds), apply the constraints.
7.2 Output the filtered stretch
8 Output segment IDs in the from/to intersections and the remaining false positive segments

**Figure 4: Sketch of the Algorithm to Chain Connected Segments**

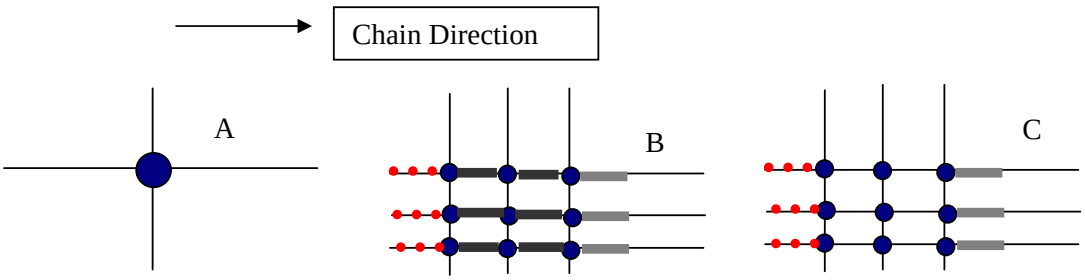


**Figure 5: Removing Segments in the From/To Intersections**

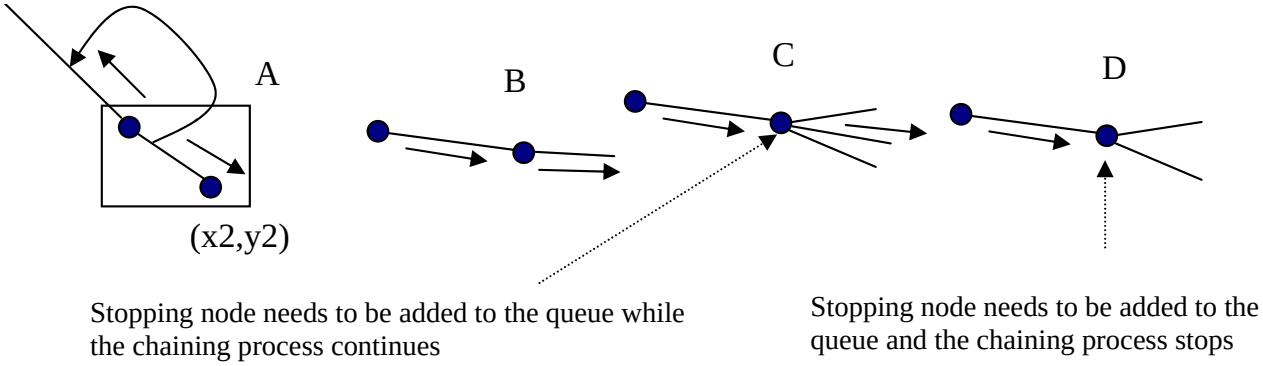


**Figure 6: Direction Computation and Making Chaining Decision**

When one of the street names in the (MainST, FromST, ToST) triple can not be mapped to a street code, automatic geocoding will not work. While some of the incorrect street names can be easily corrected (such as simple misspellings), manual interventions are required in many cases. Allowing approximate string matching may significantly increase search space if weak rules are enforced. On the other hand, the topological and geometric constraints imposed by the (MainST, FromST, ToST) triple can significantly reduce search space. We are in the process to automate this process, however, due to the level of complexity of the NYC street network and our time constraints, we have decided to follow a manual approach. Instead of correcting the incorrect street names in a traffic record individually, we also incorporate the topological and geometric constraints through a visual approach by developing a Visual Basic for Application (VBA) program on top of ESRI ArcMap software[1].

As shown in Fig. 7, when a data file contains the wrong records is loaded, the records are displayed in a list and users can select any of them to work with. The triple is then parsed into three street names and any of the streets, if it can be mapped to a street code in the LION dataset, all of its segments will be highlighted. Even if one or two street names can not be mapped, the program allows quickly zoom to the area of study defined by one or two sucessfully mapped streets and display street names as labels in the area by turning on ArcMap labeling functionality. We have observed that, when one of the mapped streets has a small number of segments, i.e., the focus area is small, other street names in the triple can be easily corrected by looking at the labels. There are also cases that all the three street names can be mapped but they do not intersect. The program allows check the topology of the three streets and users can decide whether a neighboring street should be used instead. While it is possible to send the corrected traffic record back for automatic geocoding, we have found that it is more efficient to manually identify street segments for the records, especially when only a few segments for each record are involved. The VBA tool allows select/add/delete segments for a traffic record using both the GIS map interface and a list interface (Fig. 7). The identified segments will be written out to a log file after user confirmation. A post-processing module is also developed to extract final results for the geocoding from the log file.

Our experiments have shown that, while it takes some time for a novice user to get familiar with the program and relevant ArcGIS user interface, one can achieve a manual geocode rate of a record per 1-5 minutes. The result is comparable to that reported in [10] on node-type manual geocoding although segment-type geocoding is more difficult. Among the 12,946 records in our project, as reported in the next section, 2531 were manually geocoded with high accuracy. Integrating the algorithm-based automatic geocoding and the software-assisted manual geocoding approaches has helped our project to achieve desired geocoding accuracy within budget and time constraints. In addition, the algorithms and the software are ready to be applied to similar projects which will result in a much quicker start. We next turn to the evaluation of geocoding accuracy in our project.

# 4. EVALUATION OF GEOCODING ACCURACY

In order to evaluate the accuracy of the geocoding task, the statistical Sampling Theory [9] is applied in our study. The measure for evaluation is the rate of accuracy, which is defined as the percentage of records being geocoded correctly to the total number of records in each of the datasets (volume, speed and treatment data). Since it is very time-consuming to check the correctness of geocoding for each record in the large-sized datasets (thousands of records), a sample is randomly selected from each geocoded dataset. The correctness of the geocoding is manually checked for all the records in those random samples and the rate of accuracy of geocoding for each sample is calculated. The accuracy rate for the population can then be estimated with a desired confidence level based on the statistics obtained from the sample. Here the sample and the population are used in a statistic context and refer to a subset and a whole set of the records in the original dataset, respectively. In this section, the validation process is described and the results show that our hybrid geocoding approach has achieved desired accuracy based on a rigidly designed statistic validation framework.

The validation process in our study includes three steps. The first step is to determine the appropriate sample size to meet the level of confidence and precision requirements. How many records must we randomly select to be 95% confident that population accuracy fall within the lower and upper bounds obtained from the sample? Our design allows users to specify two parameters to calculate the sample size, i.e., level of confidence ($\hat{p}$, e.g., 95%) and level of precision (e, e.g., $\pm 5\%$). The level of confidence, which is also called the margin of error, is the range in which the true value of the population is estimated to be. We follow the work of [9] and compute the sample size as $n = \frac{Z^2_{\alpha/2} \times \hat{p} \times (1-\hat{p})}{e^2}$ where n is the desired sample size, $Z_{\alpha/2} = 1.96$ for $\alpha = 0.05$, e is the desired level of precision and $\hat{p}$ is the estimated proportion of an attribute (accuracy rate) in the population. For finite population the sample size can be reduced slightly as $n' = \frac{n}{1+\frac{(n-1)}{N}}$, where n' is the adjusted sample size and N is the population size.

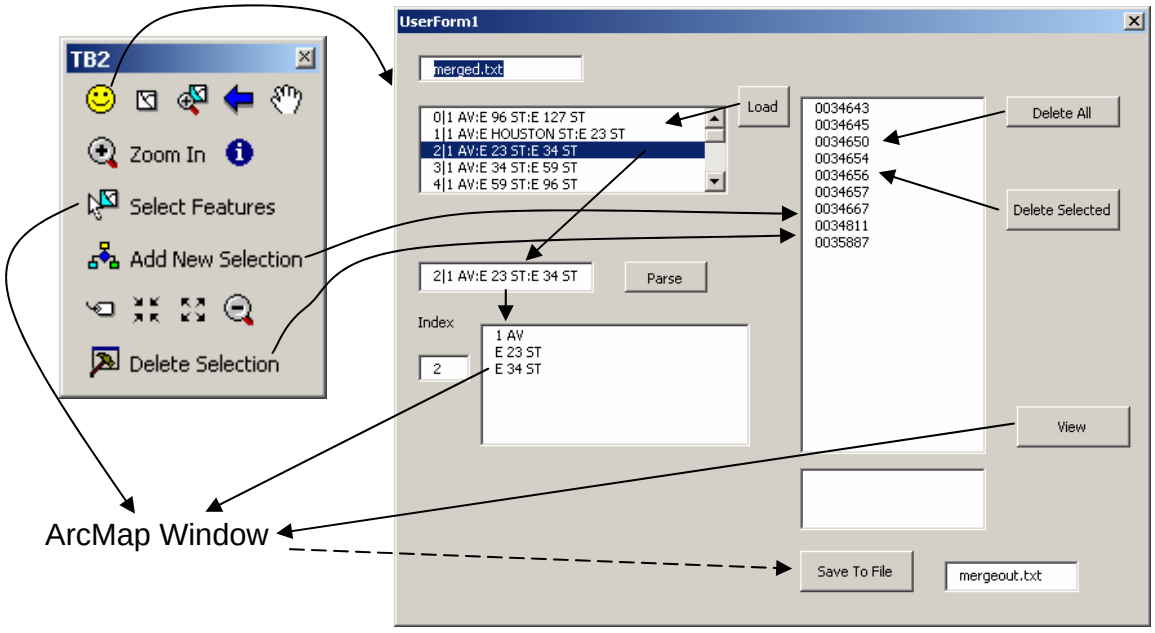


**Figure 7: Snapshot of the VBA Tool to Facilitate Maunal Geocoding**

In the second step, with the desired sample size calculated, the simple random sampling (SRS) is used to randomly choose the sample from the population set. Simple random sampling is the simplest form of the probability sampling techniques. The underlying assumption is that each record has equal likelihood to be geocoded correctly or not, i.e., error rate is independent of the characteristics of each record. The geocoding of each record in the sample is checked manually and the results (number of records geocoded correctly in each sample) are recorded.

Finally, we estimate the accuracy rate of the population. Because the actual sample size in the checking process might not be exactly the same as the calculated ones and the accuracy rate obtained from the samples might be different from the initial accuracy rate estimated from the pilot test, one can not directly apply the level of precision (such as $\pm 5\%$) from step 1 to the sample accuracy rate obtained in step 2. Instead, the confidence interval of the population accuracy rate can be estimated by $\hat{p} \pm Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$. Here $Z_{\alpha/2} = 1.96$ and n is the actual sample size used in the checking process.

The results of the accuracy rate for our geocoding of the different datasets are listed in Table 1. We can see that the accuracy rates can be considered high for all the datasets. We note that the actual accuracy rates of the geocoded datasets could be higher than the corresponding reported sample accuracy rates listed in the fourth column of table 1 for the following reasons. First, we have corrected those errors we found during the manual checking. This naturally increases the population accuracy rates. Second, we have found that many of the geocoding results are marked incorrect for long street stretches (>50 segments) even if only a small number of segments (<3) are geocoded incorrectly. However, the small percentage of incorrectly geocoded segments has very small impacts with respect to modeling accuracies. Third, we have found that most of the records that are marked as being geocoded incorrectly belong to the cases that some of the segments are missing rather than including totally irrelevant segments. Modeling accuracies are less impacted by using records with fewer segments than using records with incorrect segments.

## 5. CONCLUSIONS AND FUTURE WORK

In this study, we have reported our work on developing a hybrid approach to geocoding segment-type traffic data in NYC. The hybrid approach is a combination of algorithm-based automatic geocoding and software tool-assisted manual geocoding. While existing geocoding algorithms handle only node-type geocoding tasks, our algorithms are designed to geocode records given in the form of (MainST, FromST, ToST) triples. A VBA program has been developed on top of ESRI ArcMap to facilitate manual geocoding of problematic traffic records and those left by automatic gecoding due to the complexity level of the street network of NYC. We have also developed a rigid statistic-based framework to evaluate the accuracy of both automatic and manual geocoding results for different types of traffic data including treatment, volume and speed. Evaluations show that the geocoding accuracies vary from 88% to 98%, all with a 5% significance level.

Our work on geocoding segment-type traffic data is a first step towards supporting segment-type geocoding and we believe it has a wide range of potential applications, in addition to geocoding traffic data for modeling purposes. For our future work along the direction, first of all, we would like to refine the automatic geocoding algorithms and make them more general, efficient and robust. Second, we are interested in formalizing the geometric and topological constraints implied in a traffic record to help correct wrongly specified street names which may substantially reduce the need for manual geocoding. Third, we plan to improve our VBA tool to make it faster and more user friendly. We are considering build a standalone tool for easy distributions.

**Table 1: Results of Geocoding Accuracy Rates**

| Data types | Total records | Sample size | Sample accuracy rate | C.I. of population | Sig. Level |
|---|---|---|---|---|---|
| Treatment data | 937(auto) | 150 | 98% | (96%, 100%) | 5% |
| AADT(Volume) | 1200(auto) | 260 | 88% | (84%, 92%) | 5% |
| | 630(manual) | 51 | 94% | (87%, 100%) | 5% |
| ATR counts(Volume) | 946 (manual) | 50 | 96% | (91%, 99%) | 5% |
| Speed data | 8278(auto) | 250 | 93% | (90%, 96%) | 5% |
| | 955(manual) | 50 | 96% | (91%, 100%) | 5% |

## 6. ACKNOWLEDGMENTS

We express our gratitude to Matthew Roe at the New York City Department of Transportation, project manager on the Effectiveness of Traffic Calming study. Matthew answered our numerous questions and relentlessly obtained various datasets for us.

## 7. REFERENCES

[1] ESRI ArcGIS `http://www.esri.com/software/arcgis/index.html`

[2] NYC DCP LION `http://www.nyc.gov/html/dcp/html/bytes/dwnlion.shtml`

[3] Geosupport `http://www.nyc.gov/html/dcp/html/bytes/dwngde.shtml`

[4] MapInfo `http://www.pbinsight.com/`

[5] Wikipedia Geocoding `http://en.wikipedia.org/wiki/Geocoding`

[6] D. Ahlers and S. Boll. Retrieving address-based locations from the web, 2008. 1460015 27-34.

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

[8] F. M. Chen, R. F. Breiman, M. Farley, B. Plikaytis, K. Deaver, and M. S. Cetron. Geocoding and linking data from population-based surveillance and the us census to evaluate the impact of median household income on the epidemiology of invasive streptococcus pneumoniae infections. *American Journal of Epidemiology*, 148(12):1212–1218, 1998.

[9] W. G. Cochran. *Sampling Techniques*. John Wiley and Sons, Inc, 2nd edition edition, 1963.

[10] D. W. Goldberg. A geocoding best practices guide,. Technical report, University of Southern California GIS Research Laboratory, 2008.

[11] D. W. Goldberg, J. P. Wilson, C. A. Knoblock, B. Ritz, and M. G. Cockburn. An effective and efficient approach for manually improving geocoded data. *International Journal of Health Geographics*, 7, 2008.

[12] H. A. Karimi, M. Durcik, and W. Rasdorf. Evaluation of uncertainties associated with geocoding techniques. *Computer-Aided Civil and Infrastructure Engineering*, 19(3):170–185, 2004. Times Cited: 8.

[13] N. Krieger, P. Waterman, K. Lemieux, S. Zierler, and J. W. Hogan. On the wrong side of the tracts? evaluating the accuracy of geocoding in public health research. *American Journal of Public Health*, 91(7):1114–1116, 2001.

[14] M. D. Lieberman, H. Samet, J. Sankaranarayanan, and J. Sperling. Spatio-textual spreadsheets: geotagging via spatial coherence, 2009. 1653860 524-527.

[15] J. A. McElroy, P. L. Remington, A. Trentham-Dietz, S. A. Robert, and P. A. Newcomb. Geocoding addresses from a large population-based study: Lessons learned. *Epidemiology*, 14(4):399–407, 2003.

[16] M. Reibel. Geographic information systems and spatial data processing in demography: A review. *Population Research and Policy Review*, 26(5-6):601–618, 2007.

[17] J. C. Robinson, S. B. Wyatt, D. Hickson, D. Gwinn, F. Faruque, M. Sims, D. Sarpong, and H. A. Taylor. Methods for retrospective geocoding in population studies: The jackson heart study. *Journal of Urban Health-Bulletin of the New York Academy of Medicine*, 87(1):136–150, 2009.

[18] M. Schootman, D. A. Sterling, J. Struthers, Y. Yan, T. Laboube, B. Emo, and G. Higgs. Positional accuracy and geographic bias of four methods of geocoding in epidemiologic research. *Annals of Epidemiology*, 17(6):464–470, 2007. Times Cited: 8.

[19] T. Steenberghen, T. Dufays, I. Thomas, and B. Flahaut. Intra-urban location and clustering of road accidents using gis: a belgian example. *International Journal of Geographical Information Science*, 18(2):169–181, 2004.

[20] R. Steiner, I. Bejleri, X. Yang, and D. Kim. Improving geocoding of traffic crashes using a custom arcgis address matching application. In *22nd Environmental Systems Research Inistitute International User Confernece*, San Diego, 2003.

[21] C. Wang, X. Xie, L. Wang, Y. Lu, and W.-Y. Ma. Detecting geographic locations from web resources, 2005. 1096991 17-24.

[22] B. Yao, F. Li, M. Hadjieleftheriou, and K. Hou. Approximate string search in spatial databases, March 2010 2010.

[23] P. A. Zandbergen. A comparison of address point, parcel and street geocoding techniques. *Computers Environment and Urban Systems*, 32(3):214–232, 2008. Times Cited: 2.

[24] F. B. Zhan, J. D. Brender, I. De Lima, L. Suarez, and P. H. Langlois. Match rate and positional accuracy of two geocoding methods for epidemiologic research. *Annals of Epidemiology*, 16(11):842–849, 2006.