An Accessible and Practical Geocoding Method for Traffic Collision Record Mapping: A Quebec Case Study

Shaun Burns[1], Graduate Research Assistant
Department of Civil Engineering and Applied Mechanics, McGill University
Room 391, Macdonald Engineering Building, 817 Sherbrooke Street West
Montréal, Québec, Canada H3A 0C3
Email: shaun.burns@mail.mcgill.ca

Luis Miranda-Moreno, Assistant Professor
Department of Civil Engineering and Applied Mechanics, McGill University
Room 268, Macdonald Engineering Building, 817 Sherbrooke Street West
Montréal, Québec, Canada H3A 0C3
Phone: (514) 398-6589, Fax: (514) 398-7361
Email: luis.miranda-moreno@mcgill.ca

Joshua Stipancic, Graduate Research Assistant
Department of Civil Engineering and Applied Mechanics, McGill University
Room 391, Macdonald Engineering Building, 817 Sherbrooke Street West
Montréal, Québec, Canada H3A 0C3
Email: joshua.stipancic@mail.mcgill.ca

Nicolas Saunier, Assistant Professor
Department of civil, geological and mining engineering
École Polytechnique de Montréal, C.P. 6079, succ. Centre-Ville
Montréal, Québec, Canada H3C 3A7
Phone: (514) 340-4711 ext. 4962
Email: nicolas.saunier@polymtl.ca

Karim Ismail, Ph.D., P.Eng., Assistant professor
Department of Civil and Environmental Engineering, Carleton University
1125 Colonel By Drive, Ottawa, Ontario, Canada
Telephone: 1-613-520-2600 ext. 1709, Fax: 613-520-3951
Email: karim_ismail@carleton.ca

Word count

Text	5724
Tables (2X 250)	500
Figures (3X 250)	750
Total	6974

Date of submission: **August 1st, 2013**

[1] Corresponding Author

ABSTRACT

There have been numerous studies of geocoding systems used to assign geographical coordinates to incident reports identified simply with textual address references. These studies have typically focused on the level of accuracy achieved by various geocoding systems, and have found that acceptable results can be achieved. Depending on the quality of the input data, a match rate between 70% and 83% can be expected, with varying levels of accuracy. However, few studies have looked at the potential of freely available online geocoding services to spatially locate traffic crash records. It is proposed that although limitations currently exist, services such as the Google Maps API provide sufficient functionality and adequate accuracy for use among a wide variety of geocoding applications. A case study using traffic crash records from a municipality in the Province of Quebec is presented, with the goal of quantifying the geocoding results. It was found that although a competitive match rate is obtained, manual revision is required to ensure that the results returned by the geocoder refer to the same intersection that exists in the input address field.

INTRODUCTION

Traffic crash records are essential input data in the road safety management process. Crash data is essential in different steps of traffic safety studies and programs, including network screening (hotspot identification), safety performance function development, before-after observational studies, among others (*1*). Traditionally, collision records have been the foundation of most road safety studies and of the development of road design guides and countermeasures. In most cases in transportation engineering, police reports are the main source of collision records. Other sources, such as injury data from ambulance and hospitalization reports, are less popular. The popularity of police reports in transportation engineering, particularly within North America, is partly due to the availability of a relatively large amount of information regarding the road environment, vehicles, passengers/drivers and weather. Records typically include a number of fields, the purpose of which is to capture the consequences of the crash (number and category of all injuries, assessment of damage, etc.), the characteristics of the crash (type of impact, number of vehicles involved, environmental and roadway conditions, etc.) as well as a police officer's professional opinion on the probable cause of a crash (*2*).

Despite its acceptability, crash data suffers from weaknesses including underreporting, localization errors, varying levels of detail, missing information, and misclassification. Another important issue is the lack of accuracy in the geographical location of each crash (X-Y geographic coordinates). Although text-based address fields are included on each report, the level of detail included can create substantial ambiguity (*3*). The inclusion of accurate spatial coordinates can have important implications in the results of a traffic safety study. The miss-location of crash records can lead to wrong conclusions as the diagnosis is largely based on the relationship between traffic crashes and road network characteristics (*2*).

Geocoding methods are employed in order to link text-based addresses to X-Y geographical coordinates. In general terms, a geocoding method is defined as the process of assigning geographic coordinates to an input feature (i.e. an address) (*3*). Although satisfactory results can be obtained under ideal conditions, geocoding of crash records can suffer from a number of issues. Incomplete address input data, as well as the inclusion of shorthand notation and spelling mistakes can impact the geocoding results. Another important issue is that commercially available geocoding programs tend to exhibit a lack of flexibility when presented with different data structures and output requirements (*3*) (*4*). Additionally, programs that allow for a high degree of matching accuracy often require a large cost, increased technical knowledge, and quality reference maps upon which to base the geocoding results. There is a need for flexible geocoding methods that provide a balance between implementation complexity and the acceptable level of accuracy (*4*).

This paper proposes a simple and practical method based on online services such as the Google Maps Application Programming Interface (API) for geocoding crash data. The level of matching success as well as an indication of spatial accuracy achieved is evaluated as part of the objectives in this study. Additionally, the inherent limitations of this freely available web-based service will be discussed. In order to evaluate the accuracy of different geocoding systems and gain a better understanding of the results that can be obtained from custom API-based algorithms, a case study using crash data from the Province of Quebec, Canada is used as an application environment. The crash records used in this paper were made available through a roundabout safety project in the Province of Quebec, and include records for the period of 2000 to 2011. The data was provided by the Quebec Ministry of Transportation (MTQ) and Quebec's Automotive Insurance Board, the 'Société de l'assurance automobile du Québec' (SAAQ).

LITERATURE REVIEW

As documented in the literature, there are three main methods for the localization of crashes, depending on the location information provided: link-node or address field, route-km point and the global positioning system (GPS)-based approach. In the link-node or address approach, crash location is identified using the distance from a node, with known points along the road being identified as nodes (e.g., intersections). In some cases, the address of the event is given (street number, name and municipality, etc.). In the route-km point, one makes use of unique route numbers and unique identifiers such as mile-markers that are assigned to a continuous section of road. This is typically used for mapping highway crashes. In the third case, the coordinates of each crash record are obtained directly from GPS units at the scene of a crash, reducing errors related to the spelling, description, and transcription of address descriptors and potentially increasing the location accuracy if used correctly (*5*). Despite the advantages of a GPS-based data collection method, many jurisdictions are still reporting crash data based on the first two methods. This approach has yet to be made available to all levels of first responders, and the availability of this data cannot be relied upon. Hence, the development of crash mapping tools has been highlighted in the literature, with the basic objective to assign a location (X-Y coordinates) to each crash report, as well as taking into account potential temporal variations in location names (*3*) (*4*) (*6*).

Due to its importance in fields such as public health, police crime tracking and traffic safety research, extensive literature exists that investigates the geocoding of spatial records (*7*) (*4*) (*8*). A prevalent conclusion of the existing literature is that irrespective of the method used, geocoding results are directly related to the quality and completeness of the address input (*3*) (*4*). A number of studies have shown that electronic field-based data collection and entry can both increase efficiency and accuracy of spatial matching, due to the reduction of transcription and typographical errors (*9*). Spatial accuracy can be further improved by the collection of postal codes on all incident reports, as postal codes are typically well known and less likely to promote spelling mistakes or colloquial descriptors of the location (*10*). Nevertheless, detailed records are of little help without an appropriate geocoding algorithm to properly interpret the input data and provide an output of the desired spatial coordinates.

The typical geocoding process involves three primary steps: data standardization, record matching, and event location (*2*) (*11*) (*4*). Data standardization is an important step to consider as real-world data is known to be imperfect, with incomplete fields, incorrect formatting, misspellings, use of shorthand notation and alternative place names often being quoted as problematic (*3*) (*12*) (*10*). Record matching is generally the most important step, and can lead to the greatest error. Goldberg et al. (*13*) caution that three different types of errors can be encountered, each one having different implications on the final results. Geocoding error can be considered to come from low spatial accuracy, false matches, or from the invalidity of assumptions made during the match (*13*).

Throughout the literature, two main categories of matching algorithms can be observed; deterministic and probabilistic matching (*10*). In general, both types of matching algorithms rely on the availability of appropriate reference tables used to match a similar address to the input field and return the linked geographical coordinates. Deterministic (or rule-based) matching can be difficult to set up, with adjustments to the rules often being required (*3*). The main weakness with this type of matching is that a binary output condition exists whereby a match is either successful, or the process fails. Probabilistic matching sorts potential matches from the reference table according to the degree of separation between the input address and the reference table.

Hence, the algorithm will return the most accurate match, and provide alternative matches according to a decreasing match score (*4*).

Throughout the literature, geocoding is undertaken through both commercially available software packages and custom algorithms. These options can vary greatly in price and quality (*5*). In terms of commercial options, many of the well-known GIS programs include geocoding tools or functionality (*5*) (*3*) (*6*). Online commercial geocoding services also exist, charging either a per-request fee or through purchase of a membership (*13*). More interesting to this research, however, are those services that offer access to an application programming interface, which has the potential to provide geocoding requests free (with some limitations), or with a small fee that provides for additional functionality (*14*) (*15*). Additionally, these services provide an alternative to the rigid data structure required by typical geocoding programs.

Evaluating the accuracy of a geocoding algorithm is often a difficult task. It is widely accepted that due to the sheer volume of records that typically require geocoding, perfection would be unattainable (*12*). Accuracy is not always the most important factor, as many applications do not require high accuracy to provide meaningful results (*11*) (*4*). In public health applications, it is sufficient to assign cancer occurrences to the census-tract level (*13*). Many studies quote the percentage of matched records as a measure of geocoding performance, although match rate is fundamentally different from the accuracy measure (*5*).

In general, the literature suggests that a match rate between 70% and 83% can be considered a good rate for address geocoding (*7*) (*3*) (*9*) (*16*). Bigham et al. (*8*) state that for intersection-coded collisions a success rate of 86% is acceptable. Finding similar results, although by a different method, Ratcliffe (*12*) proposed that a minimum acceptable geocoding rate of 85% was required for the mapped data to be representative of the final map if all records had been successfully geocoded. Attempting to improve the understanding of typical geocoding error, Zandbergen (*10*) provides an estimated range of geocoding positional error to be from 25 to 168 m. With this type of inaccuracy in mind, Levine et al (*6*) suggested that although records are often provided with a directional offset from an intersection, the difficulties in interpreting the offset along with the inherent inaccuracy of the geocoded estimation render the offset value ineffective. For this reason, they mapped all incidents to the nearest intersection.

METHODOLOGY

With the goal of investigating the use of online geocoding APIs, three main steps were followed. These steps include: i) the selection of an appropriate API, ii) the development of a custom algorithm, and iii) testing of the algorithm to establish the potential match rate and accuracy that could be expected from such a service. Details of each step are explained below.

API Selection

At the beginning of this work, a number of APIs were identified and considered for further investigation. These include services offered by companies such as Yahoo!, MapQuest and Google Maps, as well as from open source services such as GISgraphy and Nominatim, which rely on OpenStreetMap geospatial data (*17*) (*18*).

A number of considerations needed to be taken into account, such as geographic coverage areas, level of detail and the reliability of finding an acceptable match given a certain quality of input data. Services that only covered the United States, for example, were not considered, as the primary dataset upon which this work is derived is from Canada. A number of tests were sent to

the various API services for reliability assessment purposes. The comparison was simplified by an API comparison tool provided on the GISgraphy website. An example of the API comparison is presented in FIGURE 1**.** As can be seen, results vary across the different services, with some being completely unacceptable.

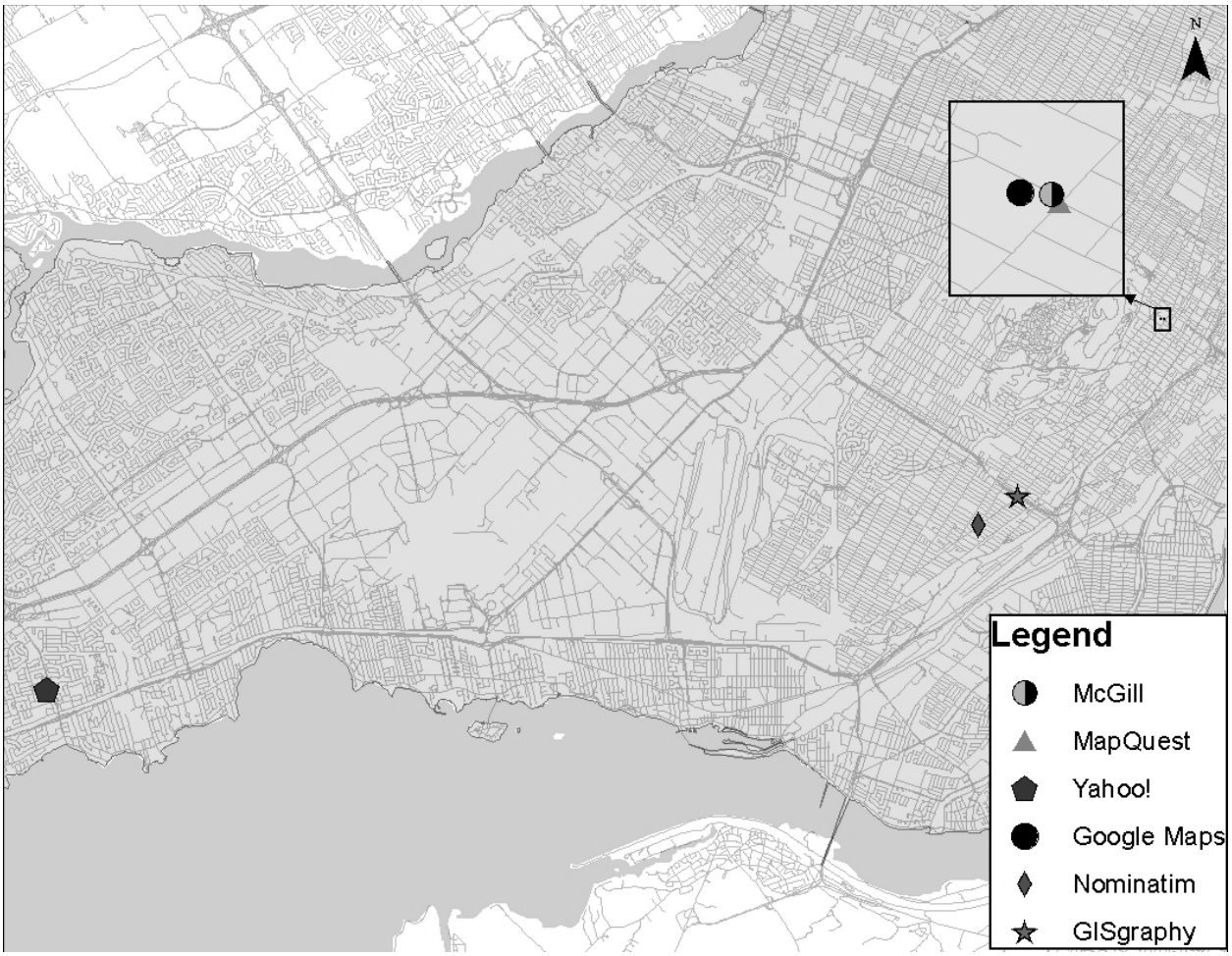

FIGURE 1 API comparison using McGill University's address (*17*).

For this paper, the Google Maps API was selected as the online geocoding service to be evaluated due to its consistency in returning addresses, and a seemingly higher accuracy (more details are presented in the results section). The API is also ideally suited for use with a mixed database of address description formats, as both link-node and address point geocoding localization methods are supported (*14*).

Programming Environment

In order to evaluate the Google Maps API, a custom algorithm was created using the Python programming language. As can be seen in FIGURE 2, the algorithm is used to read a crash database file and interpret the supplied fields. The algorithm then attempts to clean the address fields by removing redundant information (if multiple fields contain the same information, for example), and special characters (such as capital letters, accented letters (in French) which may

be misinterpreted, and punctuation). The algorithm can also be used to replace commonly misspelled words and typos, as setup by the user.

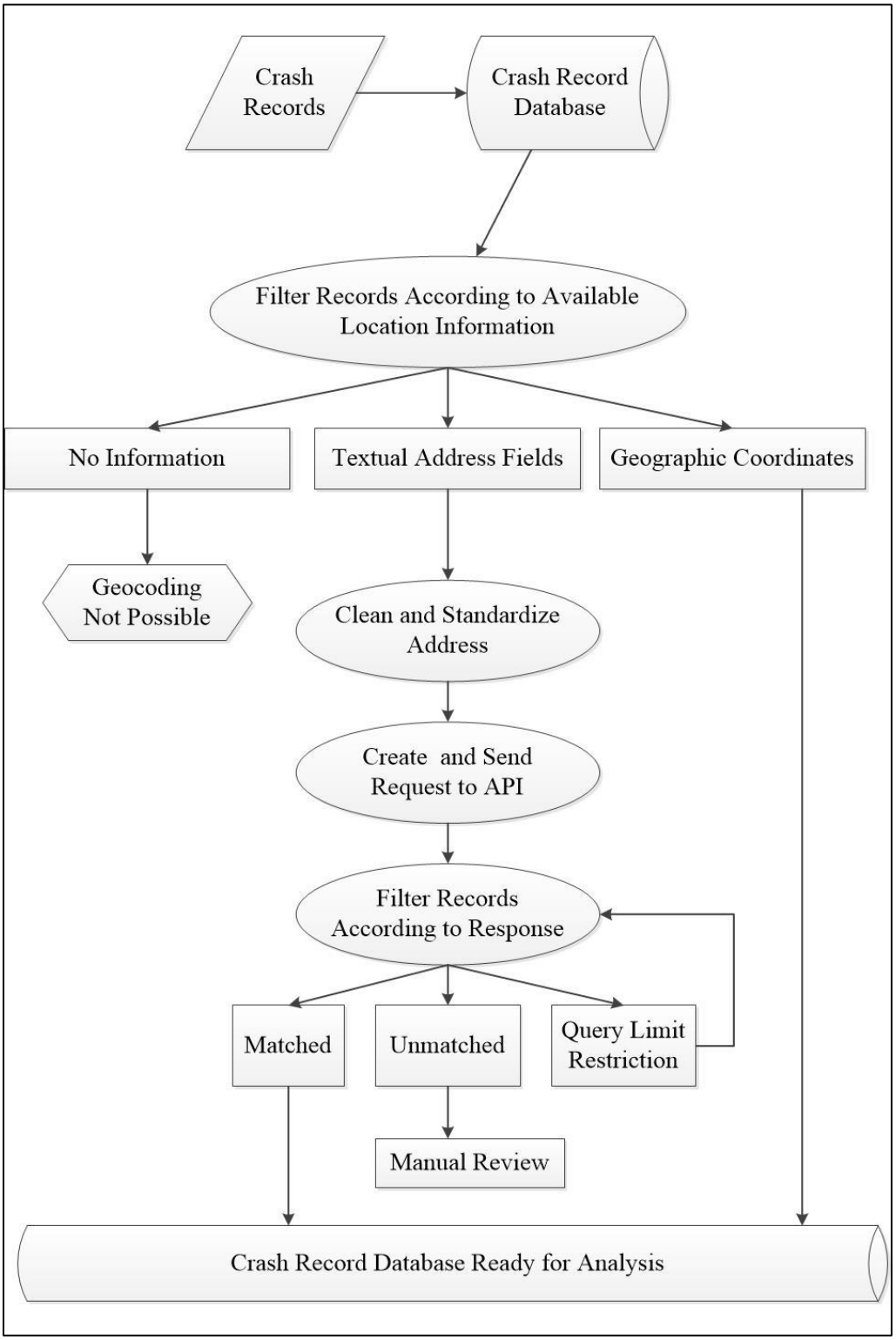

FIGURE 2 Flow chart illustrating the design of the Python algorithm.

Once this is done, the algorithm is set to call the Google Maps API using a Hypertext Transfer Protocol (HTTP) request message. An important benefit of using the API is that the address does not need to be parsed by the user. Being based on a probabilistic matching method, the API is able to parse the input, as well as infer the formatted address based on information within its proprietary reference tables. This server-side processing is beneficial as it considerably reduces the technical knowledge needed by the user.

The algorithm then collects the API's response, which can be in either *json* (JavaScript Object Notation) or XML formats (*14*). To help assess the accuracy of the returned location coordinates (i.e. latitude and longitude), the API includes a tag indicating the type of mapping accuracy that was successfully returned for a given match. The possible tags are as follows:

1. **Street_address:** Indicates that the result is a precise street address.
2. **Intersection:** Indicates that the result is at the intersection of two streets.
3. **Route:** Indicates that the result is a named street segment.
4. **Political, country, administrative_area_level_1, administrative_area_level_2, administrative_area_level_3, locality, sublocality, neighborhood:** Indicate that the result is within a political or civil entity (such as a municipality, province, etc.)
5. **Colloquial_area, premise, subpremise:** Indicates that the result is a named location, such as a well-known building.
6. **Postal_code:** Indicates that the result is a postal area.
7. **Natural_feature, airport, park:** Results are as indicated.
8. **Point_of_interest:** Indicates that the result is a local point of interest that does not fit in another category (*14*).

The API's responses are returned and ranked from the most to the least accurate match levels. For this paper, only the first three tags (i.e. *street_address, intersection* and *route*) were considered to be useful matches; all subsequent tag levels were considered to have not returned a match.

Analysis and Mapping

In order to evaluate the quality of the geocoding achieved by the Google Maps API implementation, two quality measures were analyzed. The first is the match rate. As mentioned above, a match is achieved if *street_address*, *intersection* or *route* was returned as a match indicator tag. The *route* tag is less accurate than the other two options due to the fact that it indicates that the record occurred somewhere along a street segment. Nevertheless, it is considered to be of sufficient accuracy for intersection safety analysis studies as it is still possible to conclude to which route the crash belongs (and to exclude it from analysis at intersections).

The second quality measure attempted to capture the level of accuracy that was provided by the geocoder. This was done by comparing the latitude and longitude provided by the Google Maps API geocoder to those that were already provided with some of the Quebec crash records. Although a comparison to the actual location as found on a map would be more representative of the true error, this would be impractical due to the number of records contained in the dataset. The results of the comparison with the previously geocoded records are presented below.

DATA

The data used as a test case for calling the Google Maps API geocoder was obtained from the MTQ as well as the SAAQ as part of a roundabout safety study being conducted by the authors. The data was presented as a digitized collection of traffic crash reports originally filled in by law enforcement officials at the scene of a crash between the years 2000 and 2011.

To evaluate the advantages and accuracy of the methodology, records from the municipality of Amos, Quebec were considered. This municipality was selected due to the large proportion of non-geocoded crash records, as well as its remote nature. The logic behind this was that if the API is capable of geocoding a smaller, remote municipality, it should be able to handle larger municipalities as well.

The total number of crash records for this municipality is 6641 records, with only 50% (or 3322) of records having been supplied with coordinate references. It is interesting to note that of the 3319 crash records that lack coordinates, only 22 records are from crashes occurring on roadways under provincial (MTQ) jurisdiction. This is most likely due to the addition of geographic coordinates at the time of digitization of the records.

Although a number of fields are contained in each record relevant to the crash, the primary focus in this study is on the location fields. These include:

- **ADR_NUMR_IMMBL:** Street number of a house/building near the crash site.
- **ADR_NOM_VOIE:** Street name on which the crash occurred.
- **VAL_NUMR_ROUTE:** Route number if applicable (such as a numbered highway, etc.).
- **NOM_VOIE_INTSC:** The name of a cross-street if the crash occurred at, or in proximity to an intersection.
- **VAL_AUTRE_IDENT_REPR:** Name of other identifying landmark if available.
- **VAL_DISTN_REPR:** Distance (in metres) to the intersection or landmark.
- **DES_TYPE_DIRCT:** Direction from the crash (if distance is not 0).

As with any form of real-world data, these fields are not always filled-in correctly. Because of this, it is possible to observe records with incomplete address information. Common issues include missing street numbers, partial street names, lack of a cross street or other landmark, among others. Each of these issues leads to a reduction in the address quality, and reduces the chances of accurate location information being returned by a geocoder.

RESULTS & DISCUSSION

API Selection

As previously mentioned, a number of APIs were considered for use in a custom algorithm. In order to determine which API was most likely to reliably return responses with an acceptable accuracy, the output was compared for a number of test cases. TABLE 1 presents a sample of these test cases, as well as an indicator to outline whether the coordinates are valid for the input location.

As can be seen, very different performance is obtained from the APIs. Surprisingly, Nominatim returned no results for any of the input addresses. It is possible that the API is not able to interpret the addresses that are being sent to it, or that its reference tables do not cover the region in question. Therefore, this API was dropped from consideration, as it would not allow for

the geocoding of the available dataset. GISgraphy was similarly dropped due to the fact that the results returned were so inaccurate that they did not even fall in the municipality of interest. Both Yahoo! and MapQuest performed similarly, with half of the input addresses being returned successfully. The Google Maps API was ultimately selected for further analysis in this paper as it returned valid coordinates for all but the last test case. The last test case was handled in the same way by all of the proprietary source geocoders: due to the incomplete nature of the address, a guess was made as to the full address.

TABLE 1 Detailed API comparison for (A) APIs with proprietary data and (B) open source data.

(A) Proprietary Source

Address (in Amos, Quebec)	Google Maps			Yahoo!			MapQuest		
	Longitude	Latitude	Valid?	Longitude	Latitude	Valid?	Longitude	Latitude	Valid?
Des Metiers at Av Du Parc	-78.1229	48.5608	Yes	-78.1231	48.5607	Yes	-78.1231	48.5607	No
343 6e Rue Ouest	-78.1309	48.5693	Yes	-78.0121	48.6110	No	-78.1311	48.5686	Yes
94 Principale Sud at Du Metro	-78.1158	48.5697	Yes	-78.0121	48.6110	No	-78.1160	48.5736	No
4e Rue Est at Gravel	-78.1063	48.5650	Yes	-78.1065	48.5649	Yes	-78.1065	48.5649	Yes
82 1e	-78.1330	48.5731	Guess	-78.1133	48.5719	Guess	-78.1176	48.5719	Guess

(B) Open Source

Address (in Amos, Quebec)	GISgraphy			Nominatim		
	Longitude	Latitude	Valid?	Longitude	Latitude	Valid?
Des Metiers at Av Du Parc	-73.7058	45.5531	No	-	-	No
343 6e Rue Ouest	-73.8667	45.5480	No	-	-	No
94 Principale Sud at Du Metro	-73.3233	45.3214	No	-	-	No
4e Rue Est at Gravel	-73.6299	45.6001	No	-	-	No
82 1e	78.1064	48.5659	No	-	-	No

*Validity refers to whether the returned coordinates are within an acceptable distance from the true coordinates of the address.

Matching Proficiency

The algorithm output was obtained in a comma-separated file that could be analyzed in Microsoft Excel software to establish a preliminary match rate and accuracy estimation. For the Amos, Quebec crash records, it was found that of the 3319 records that lacked geographic coordinates, 2586 (or 78%) of records were matched to either an *intersection* or *street_address* level. Assuming that the results are being used to perform an intersection safety analysis, it would also be possible to include the *route* results, as this would locate the records along a given route, indicating that they did not occur at an intersection. With this assumption, the match rate is found to increase to 85%. Adding these records to those that were previously geocoded by either the MTQ or the SAAQ, it is found that over 92% of all traffic crash records can be mapped for the municipality. Similar results were obtained for tests performed on datasets from other municipalities in the Province of Quebec, although the results are not presented here.

From the results, it can be seen that using a custom algorithm to call upon an online geocoder service can provide a competitive match rate that falls within the accepted rate in the literature for commercially available systems. The main benefit of this method, however, is that it is not required to parse the input address information to match a specific format before passing the input to the geocoder. This information is automatically extracted by the geocoder, with a seemingly high level of confidence.

Accuracy Estimation

In terms of the accuracy estimation, results are less conclusive. As documented in the literature, measuring the accuracy is a difficult task to undertake, and often requires manual verification in order to obtain any level of confidence in the conclusion.

As previously mentioned, the results of the accuracy estimation were obtained by comparing the distance between the known coordinates and those provided by the algorithm. Looking at the raw results, a large discrepancy could be observed for a number of entries. A more comprehensive analysis found that the Google Maps API handled the geocoding of records identified by route number alone (and not the more common name of the road segment) very poorly. Removing these records from the estimation it was found that the average distance between the known and geocoded coordinates is 200 m. An interesting observation however, is that 54 % of the records have a distance between the two coordinate estimations of less than 30 m.

Initially, the average distance error between the previously geocoded coordinates and those obtained with the use of the custom algorithm seems relatively high at 200 m. Looking at the data, however, it can be seen that a wide range of estimations is obtained. Selecting a sample of records with a higher degree of match quality, however, yields an estimated distance of only 22 m.

One consideration that was investigated in order to clarify this result is that the previously geocoded records were taken to be accurate representations of the crash location, although it is possible that they are in fact estimations in and of themselves. Looking at FIGURE 3, this hypothesis seems to be a possibility, as neither the previously geocoded coordinates, nor those obtained from the algorithm are at the true location indicated in the crash record. A manual sampling of the results reveals that in fact, the records from the algorithm are more accurate than the previously known coordinates in many cases. From this, it can be concluded that estimating an accuracy measure by comparing the results to those previously geocoded may be flawed, and that the Google Maps API may in fact provide better estimates than originally thought provided a high quality address record is available.

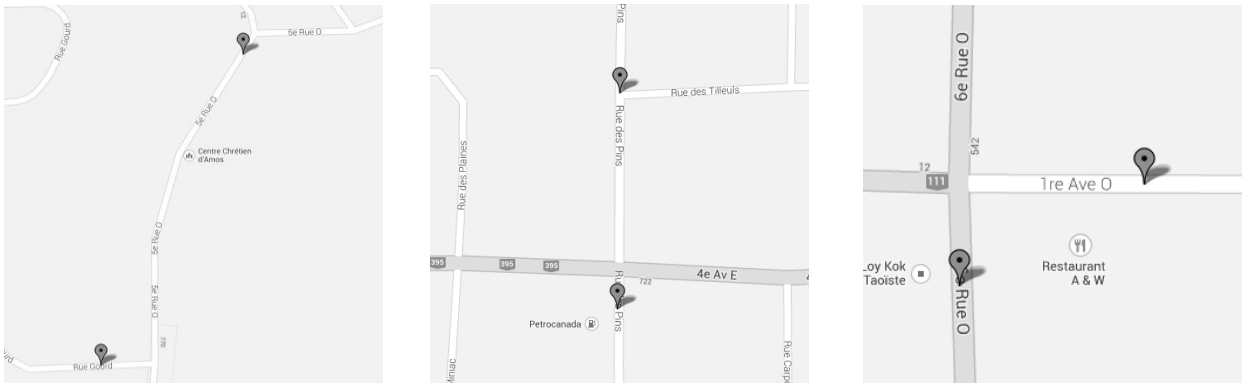

FIGURE 3 Examples of coordinate differences (*14*).

One of the main observations that should be taken away from the preliminary results, however, is that the Google Maps API tends to always provide a match for a given input, even if the match quality and accuracy are low. This may lie in the fact that the user has no direct control on the probabilistic matching limits, and thereby is forced to accept the result returned by the API. Because of this, revision of the resulting matches is suggested, as is caution in the use of the returned results. As manual revision is impractical for large datasets, an automated process should be investigated in order to improve the reliability of the geocoding.

Common Causes of Low-Quality Matching

Looking at the records with a high estimation of distance between the known and geocoded coordinates reveals that a majority of these records have some sort of ambiguity involved in their address fields. This ambiguity prevents the geocoder from returning high quality results. A summary of common shortcomings is presented in TABLE 2 below.

TABLE 2 Sample Address Problems Causing Low-Quality Matches.

Address	Problem
622 Des Javies, Amos, Quebec	Street does not exist in the municipality.
Ruelle Arriere Restaurant Succo, Amos, Quebec	This description is not recognized by the geocoder.
1132 RTE 111 E at 4e rue E, Amos, Quebec	The geocoder has difficulty identifying numbered roadways.
Taschereau at 10 Av E, Amos, Quebec	The geocoder interprets the "10" as a house number, and not the street number due to lacking formatting (i.e. 10e av.).
Av Authier O, Amos, Quebec	No street number is provided as reference on the street, so a general segment location of Avenue Authier is returned.
1e at 2e, Amos, Quebec	The street type is missing both cases, leading to a guess on the location.
22 Principale, Amos, Quebec	No distinction between Principale North or South. Although a match is returned, the API provides a guess as to which street is meant.

Examples of the address record shortcomings include records located on streets with both "North" and "South" components with no distinction provided in the record, as well as records with addresses such as "1e at 2e". Without the inclusion of street types it is difficult for the geocoder to determine if this is a cross of First Street and Second Avenue, or a similar combination.

Limitations

Among the limitations of the proposed algorithm, readers should take note that at present the APIs presented in this paper are intended for the use of online application developers to include maps on their respective websites and/or mobile apps. Any use beyond this requires special permissions be obtained from the API owners. As such, this work remains largely a proof of concept, with the purpose of showing the potential applications of the technology that is currently on the market.

CONCLUSIONS AND FURTHER WORK

This study has explored the use of online geocoding services such as the Google Maps API as a simple and accessible tool for the geocoding of traffic crash records. It was found that at the strictest level, a match rate of 78% could be achieved through a custom algorithm. Relaxing of the matching conditions improved the match rate to 85%, although caution should be taken as not all applications can support the associated reduction in match reliability provided. These results are comparable to those obtained from commercially available geocoding options, although manual review indicated that a number of false matches occurred when incomplete input data was sent to the geocoding API.

The study also explored the geocoder's spatial accuracy, although the results tend to vary substantially from record to record. Factors such as the completeness of the input address fields and the ability of the API to interpret the location description (in particular for route number addresses) have a large influence on this outcome.

It is suggested that with proper user revision, the results are sufficient for practical applications such as intersection safety analysis. The use of an intersection's area of influence should compensate for at least some of the observed inaccuracies, and allow for the crash records to be associated with either the respective intersection location, or street segment to which they belong.

Future work will focus on improving the algorithm in such a way as to introduce greater error checking capabilities. Work will also focus on obtaining the required licensing to expand the sample records tested with the geocoding API in order to obtain an improved accuracy estimation measure. Finally, the possibility of testing other internet-based geocoders will also be investigated.

ACKNOWLEDGEMENTS

The authors would like to thank the Québec Ministry of Transportation (MTQ) as well as Quebec's Automotive Insurance Board (SAAQ) for providing the data used in this paper. Funding for the roundabout research project was provided by the Québec Fund for Research on Nature and Technology (FQRNT), the Québec Ministry of Transportation (MTQ), and the Québec Fund for Health Research (FRSQ) as part of their research program on road safety.

REFERENCES

1. National Research Council (US). Transportation Research Board. Task Force on Development of the Highway Safety Manual and American Association of State Highway and Transportation Officials. *Highway Safety Manual*. AASHTO, Washington D.C., 2010.
2. Chen, J. Black Spot Determination of Traffic Accident Locations and Its Spatial Association Characteristic Analysis Based on GIS. *Journal of Geographic Information System*, Vol. 4, 2012, pp. 608-617.
3. Christen, P., T. Churches, and A. Willmore. A Probabilistic Geocoding System based on a National Address File. in *The Australasian Data Mining Conference*, Cairns, Australia, 2004.
4. Davis Jr., C. A., F. T. Fonseca, and K. A.d.V. Borges. A flexible Addressing System for Approximate Geocoding. in *Proceedings of the Fifth Brazilian Symposium on geoInformatics*, Sao Paulo, Brazil, 2003.
5. Goldberg, D. A. A Geocoding Best Practices Guide. University of Southern California, Guide, 2008.
6. Levine, N., and K. E. Kim. The Location of Motor Vehicle Crashes In Honolulu: A Methodology For Geocoding Intersections. *Computers, Envrionment and Urban Systems*, Vol. 22, no. 6, 1998, pp. 557-576.
7. Velavan, K. Developing Tools and Data Model for Managing and Analysing Traffic Accident. University of Texas at Dallas, Dallas, Texas, Thesis 2006.
8. Bigham, J. M., 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*, Vol. 8, no. 72, December 2009.
9. Cherry, E., R. Floyd, T. Graves, S. Martin, and D. Ward. Crash Data Collection and Analysis System. ARCADIS G&M of North Carolina, Inc. , Pheonix, Arizona, Final Report FHWA-AZ-06-537, 2006.
10. Zandbergen, P. A. A Comparison of Address Point, Parcel and Street Geocoding Techniques. *Computers, Environment and Urban Systems* , Vol. 32, 2008, pp. 214-232.
11. Davis Jr., C. A., and F. T. Fonseca. Assessing the Certainty of Locations Produced by an Address Geocoding System. *Geoinformatica*, Vol. 11, 2007, pp. 103-129.
12. Ratcliffe, J. H. Geocoding Crime and a First Estimate of a Minimum Acceptable Hit Rate. *Juornal of Geographical Information Science*, Vol. 18, no. 1, January-February 2004, pp. 61-72.
13. Goldberg, D. W., J. P. Wilson, and C. A. Knoblock. From Text to Geographic Coordinates: The Current State of Geocoding. *URISA Journal*, Vol. 19, no. 1, 2007, pp. 33-46.
14. Google, Inc. The Google Geocoding API. *Google Developers*, 2013. https://developers.google.com/maps/documentation/geocoding/. Accessed July 25, 2013.
15. Yahoo!, Inc.. Yahoo! Maps Web Service. 2013. http://developer.yahoo.com/maps/. Accessed July 25, 2013.
16. Trahan, S., M. Nguyen, I. Allred, and P. Jayaram. Integrating Geocode Data from the Google Map API and SAS/Graph®. in *Preceedings of SouthEast SAS Users Group (SESUG) Conference*, North Carolina, 2009.
17. Qin, X., S. Parker, Y. Liu, A. J. Grettinger, and S. Forde. Intelligent Geocoding System to Locate Traffic Crashes. *Accident Analysis and Prevention*, Vol. 50, 2013, pp. 1034-1041.
18. GISgraphy. *GISgraphy Results Comparator*, 2012. http://www.gisgraphy.com/compare/. Accessed July 28, 2013.

19. OpenStreetMap. OpenStreetMap. 2013. http://www.openstreetmap.org/. Accessed July 28, 2013.

20. Steiner, R., I. Bejleri, X. Yang, and D.-H. Kim. Improving Geocoding of Traffic Crashes Using a Custom ArcGIS Address Matching Application. in *22nd Environmental Systems Research Institute International User Conference*, 2003.