Transportation Research Part C 68 (2016) 185–193

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Extracting accurate location information from a highly inaccurate traffic accident dataset: A methodology based on a string matching technique

Mario Miler [a,*], Filip Todić [a], Marko Ševrović [b]

[a] Department of Geoinformatics, Faculty of Geodesy, University of Zagreb, Kačićeva 26, 10000 Zagreb, Croatia
[b] Department of Transport Planning, Faculty of Transport and Traffic Sciences, University of Zagreb, Vukelićeva ulica 4, 10000 Zagreb, Croatia

ARTICLE INFO

Article history:
Received 1 April 2015
Received in revised form 21 September 2015
Accepted 4 April 2016

Keywords:
Traffic accident
OpenStreetMap
Jaro–Winkler
Inverse Distance Weighting
Data validation

ABSTRACT

The objective of this research was to develop a model for validating traffic accident locations that would be applicable worldwide, regardless of linguistic or cultural differences. In order to achieve this, a Volunteered Geographic Information (VGI) dataset was used, the OpenStreetMap (OSM) project. To test the developed model, a total of 8550 accidents with fatal or non-fatal injuries that occurred in the City of Zagreb from 2010 to 2014 were evaluated. Traffic accident data was collected using the pen-and-paper method while the traffic accident locations were determined using Global Positioning System (GPS) receivers embedded within police vehicles. This form of data entry invariably introduces errors in both geometric and contextual attributes. To fully counteract these errors, the developed model consists of two key concepts: the Jaro–Winkler string matching technique and the Inverse Distance Weighting method. Over 66% of traffic accident locations were validated, which is an increase of 15% when compared to the classical approach. The model outlined in this paper shows a significant improvement in estimating the correct location of traffic accidents. This in turn results in a drastic decrease in resources needed to estimate the quality of accident locations.

1. Introduction

According to the World Health Organization (2013), road traffic injuries are the eighth leading cause of death globally and considering the current trends, by 2030 they will become the fifth leading cause of death. Road accidents are a direct result of the increased mobility of contemporary society. As noted by Anderson (2009), their impact in terms of injuries and fatalities is a social and public health challenge.

Loo (2006) has argued that while a highly sophisticated statistical and mathematical models can be built to prevent future occurrences of accidents, the integrity of the results depends heavily on the availability and quality of collected accident data. Most cities and countries rely on the police force for collecting, storing and publishing traffic accident data. The traffic accident data available to us was collected primarily for administrative purposes rather than scientific analysis. Although a high level of precision about the location of a traffic accident may not be vital for all accident analyses, it is

* Corresponding author.
E-mail addresses: mmiler@geof.hr (M. Miler), fitodic@geof.hr (F. Todić), sevrovic@fpz.hr (M. Ševrović).

http://dx.doi.org/10.1016/j.trc.2016.04.003
0968-090X/

essential to know the accuracy of any given dataset. This is crucial for any meaningful spatial analysis ranging from simple visualizations of spatial patterns to more complex analyses of underlying spatial trends.

As noted by Tegge and Ouyang (2009), most of the safety analysis models are built upon spatially and temporally matched traffic accident and roadway datasets, errors and inconsistencies in traffic accident location records are a realistic problem that often compromises accuracy of safety model outcomes. These problems may occur for numerous reasons (a) traffic accident location information is generally extracted from paper-based police reports which accuracy highly depends on the skill and experience of the police officers on site, as well as the motivation of other persons involved in data processing, (b) traffic accident information and road geometry datasets are often developed on different software platforms that utilize different spatial coordinate systems and formats (evident upon merging datasets), (c) most accidents are located at the borderline of two or more different road segments because such locations are easily identifiable by police officers.

Khan et al. (2004) reviewed international practices and technologies used in accident data management. At the time, they found that the pen-and-paper method was the most commonly used method for recording accident data in the field. Loo (2006) concluded that most pen-and-paper forms found to be 2–4 pages long and were filled under difficult circumstances which prohibits police officers from making detailed and accurate records of all relevant data. However, despite the recent advancements in technology, Burns et al. (2014) noted that these practices had not improved.

Traffic accident locations have to be evaluated before conducting any meaningful spatial analysis and our model should be able to decrease the resources needed to evaluate the quality of these locations and identify the ones that are likely to be incorrect. Since the model was designed to evaluate traffic accident locations, a digital database containing traffic accident records was needed. These records should include geographic coordinates of all accidents and, if possible, references to cities, streets or roads where they occurred. The accident dataset used in this study had unreliable location data (latitude and longitude) due to several factors, including the use of various Web services and the misuse of GPS devices. The dataset also contained various street name abbreviations as a direct result of using the pen-and-paper method. When compared to the accident datasets used in the existing literature, the quality of the accident dataset, including the availability of certain attributes is much poorer.

In order to create a universally applicable model, a publicly accessible dataset was needed. One of the these publicly available datasets is the OpenStreetMap (OSM) project. Mooney and Corcoran (2012) definded the OSM project as a collaborative project which aims at creating a free editable map (road) database of the world and it is the most well known example of Volunteered Geographic Information (VGI) system. VGI refers to the collaborative collection of geographic information by citizens.

Haklay (2010) and Neis et al. (2011) analyzed the quality of OSM dataset and indicated that the quality of the OSM dataset is on the rise. Therefore, one of the objectives of our research was to evaluate its potential for researches in traffic safety. This research employs a string matching technique to identify the most probable street for each accident. String matching techniques have already been used on the OSM dataset by Mooney and Corcoran (2012) for detecting changes in attribute data on heavily edited objects. In our research, the same technique was used to compare the street names registered in police records to the street names from the OSM. Considering the possible discrepancies between the records in the accident and the OSM databases, one of the main challenges was to estimate at what level the string matching algorithm was suitable for determining the location of a traffic accident based on the street names registered by the police.

When comparing street names between accident and the street network datasets, exact name matches are the most preferable result. However, there is no guidance in the existing literature as to the most suitable lower value of similarity percentage between two street names. Therefore, several suggestions regarding acceptable similarity percentages will be proposed in this paper. The objective of our research was to develop a model capable of extracting accurate traffic accident locations from a highly inaccurate dataset that would be applicable worldwide, regardless of linguistic or cultural differences.

2. Methods

The model described in this paper was the result of a project commissioned by the Ministry of Interior of the Republic of Croatia. It was successfully implemented on a national scale and produced outstanding results. However, in this paper only 8550 fatal and non-fatal accidents located in the City of Zagreb were analyzed in detail.

The developed model was implemented in the Python programming language. It also includes the use of the Jaro–Winkler string matching technique as described by Bilenko et al. (2003) and the Inverse Distance Weighting method as described by Bakkali and Amrani (2008).

The study area in this research was the City of Zagreb, the largest urban metropolitan area in Croatia. The traffic accident dataset for the study area was obtained from the Ministry of Interior of the Republic of Croatia in a *Comma Separated Value* (CSV) file. The dataset comprises all accidents that occurred from 2010 to 2014 inclusive. A total of 8550 fatal or non-fatal accidents were analyzed in detail. The street network dataset used in this research was obtained from the OpenStreetMap (OSM) project on February 1st, 2015.

After acquiring the necessary datasets, the data was imported into a PostgreSQL database which was extended with PostGIS, an open-source spatial database extension for the PostgreSQL object-relational database.