ELSEVIER

Accident Analysis and Prevention 38 (2006) 879–886

ACCIDENT ANALYSIS & PREVENTION

www.elsevier.com/locate/aap

Validating crash locations for quantitative spatial analysis: A GIS-based approach

Becky P.Y. Loo *

Department of Geography, The University of Hong Kong, Pokfulam, Hong Kong

Received 23 November 2005; received in revised form 24 February 2006; accepted 24 February 2006

Abstract

In this paper, the spatial variables of the crash database in Hong Kong from 1993 to 2004 are validated. The proposed spatial data validation system makes use of three databases (the crash, road network and district board databases) and relies on GIS to carry out most of the validation steps so that the human resource required for manually checking the accuracy of the spatial data can be enormously reduced. With the GIS-based spatial data validation system, it was found that about 65–80% of the police crash records from 1993 to 2004 had correct road names and district board information. In 2004, the police crash database contained about 12.7% mistakes for road names and 9.7% mistakes for district boards. The situation was broadly comparable to the United Kingdom. However, the results also suggest that safety researchers should carefully validate spatial data in the crash database before scientific analysis.

Keywords: Police-collected crash information; Spatial variables; Database; Geographic information system (GIS); Spatial data validation; Spatial analysis

1. Introduction

How useful is accident analysis in preventing the future occurrence of accidents? This is a fundamental question to people interested in *Accident Analysis and Prevention*. While highly sophisticated statistical and mathematical models can be built, the validity of modelling results still lies critically on the availability and quality of accident data. This paper proposes a methodology that validates and identifies the precise road crash location with the link-node system with no buffer zone. Moreover, it goes beyond reporting the extent of mis-location or mis-identification in the crash database. Spatial variables, if found to be mis-coded, are corrected for further spatial analysis. The methodology is applied to examine all police-reported road crashes in Hong Kong from 1993 to 2004. In this way, temporal changes in the quality of the spatial variables in the crash database can be identified. While the Hong Kong road crash database has it own specific structure, the methodology and findings of this paper are of general interest to road safety researchers elsewhere.

In most cities and countries, the primary responsibility of collecting road crash data rests with the police. The major aim of collecting the crash information by the police, however, is primarily for administrative purpose (including litigation but also monitoring, evaluation, and problem detection) rather than for scientific analysis. Recently, Khan et al. (2004) closely examined the road crash report forms used by the police in various administrations, including Australia, Emirate of Dubai (United Arab Emirates), the Kingdom of Bahrain, New Zealand, Sweden, the United Kingdom and several states in the United States of America. They found that these report forms could yield up to 99 pieces of information in relation to the general information, location, road users, injury details, road environment, vehicle(s) and crash characteristics. However, is the quality of these data good enough for accident analysis and prevention?

Over time, many researchers (Austin, 1995; Ibrahim and Silcock, 1992; Shinar and Treat, 1979; Shinar et al., 1983) have raised questions about the accuracy, precision and reliability of road crash information collected by the police. The reason was partly due to the lengthy road crash report forms, often 2–4 pages long, which require filling-in at the crash sites, mostly by the pen-and-paper method. Often, the circumstances (with causalities, other emergency service personnel like firemen and ambulance men, and impatient/curious road users affected by

* Tel.: +852 2859 7024; fax: +852 2559 8994.
E-mail address: bpyloo@hkucc.hku.hk.

0001-4575/$ – see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aap.2006.02.012

the traffic disruption) prohibit the police officers from making detailed and accurate records of all relevant data. Furthermore, the police cannot be regarded as professionals for all information, such as vehicle defects, drivers' state and conditions, and environmental deficiencies (Shinar and Treat, 1979).

Consequently, the quality of police-collected crash data is not uniform for all variables. In particular, spatial variables were found to be the *most* reliable in the study of Shinar et al. (1983), whereby a multi-disciplinary accident investigation (MDAI) team made independent assessments in relation to the crash characteristics, vehicle characteristics, driver characteristics and crash causes. The MDAI assessments on a sample of 124 crashes were then compared with the police records to identify discrepancies. It was found that the most reliably reported data were crash location, date, and number of drivers, passengers, and vehicles. In sharp contrast, crash location was found to be the *least* reliable in the studies of Ibrahim and Silcock (1992), Austin (1993) and Khan et al. (2004). Based on a representative survey of Highway Authorities in Great Britain (73 out of 93), Ibrahim and Silcock (1992) found that "the problem which occurred most frequently is the inaccuracy of the accident location by the grid reference" (p. 494). Nearly half of the Highways Authorities reported that inaccurate crash location by the grid reference was their most frequently encountered problem. Similarly, the study of Austin (1993) found that "accident location was considered to be the most incorrectly coded" police-reported variable (p. 540) and the problem was more serious than previously reported. Similarly, Khan et al. (2004) remarked that "the single biggest problem with the quality of accident data in Abu Dhabi has been the disregard to identify and record the precise location of the accident" (p. 2).

While a high level of precision about the location of a road crash may not be central for all crash analysis, it is essential for any meaningful spatial analysis, which ranges from the simple visualization of spatial patterns and the identification of hot spots to the more complex analyses of underlying spatial trends, spatial interactions and spatial autocorrelation. By and large, all spatial statistics, such as centrography (like spatial mean and median), standardized nearest neighbour index, variance–mean ratio and crash density, are affected by the precision of spatial data. The requirement for precision is the greatest when the spatial unit of analysis is small, such as in hot spot analysis and network autocorrelation analysis. Thus, it is vital to assess and validate the spatial location of road crashes before conducting scientific spatial analysis.

In this paper, a spatial data validation system is developed using the geographic information system (GIS). Generally, the aim is two-fold. Firstly, it aims to provide estimates about the levels of accuracy, precision and reliability of the spatial variables of road crashes collected by the police. Secondly, it aims to improve the raw road crash database by identifying and correcting mis-specifications in the spatial variables for scientific analysis. The rest of the paper is organized as follows. The next section provides a review of previous attempts to code and/or validate police-collected spatial variables. The ways how the present paper differs from previous studies are highlighted. The GIS-based spatial data validation system is then introduced. Lastly, results of the validation are presented and discussed. The paper concludes by highlighting the implications of spatial data validation and the ways forward for improving the quality of crash information.

2. Literature review

Conceptually, Kam (2003) proposes several alternatives for geocoding road crashes. They are exact address geocoding, approximate geocoding and map grid geocoding. The first alternative is the most preferable if the precise location of a crash, such as the exact address, is identifiable. The second alternative adopts the methodology of randomly selecting a point along a given road when only the road name is known. The third alternative resorts to the map grid and randomly selects a point within the map grid to represent the crash location. This method relies on an internal program to judge, validate and locate the site of a crash within "permissible" locations, such as along the road network. For a spatial analysis of crash patterns to yield meaningful results, exact address geocoding is the most preferable. Given the focus of Kam (2003)'s paper on crash rate analysis, it was only reported that "accident locations are geocoded based on the horizontal and vertical grid reference to Melway" (p. 699). No pre-analysis data validation has taken place.

In contrast, Levine et al. (1995) demonstrated the procedures for geo-referencing crash locations in practice before analyzing the degree of spatial concentration. In their study of Honolulu, Hawaii, all road crashes were snapped to the nearest road intersections/junctions before the spatial analysis. Levine et al. (1995) first developed a standardized dictionary of street names by using the "AutoStan" software. The street names were then matched with the files of topologically integrated geographic encoding and referencing (TIGER). A set of alternative street names was provided for the dictionary in order to derive a high proportion of successful matching for the road crashes. Afterwards, the intersections of road segments were given the street names for the intersected streets, and the latitude and longitude of the intersections were compared with those of the street names. If the matching was successful, the intersection was assigned as the crash location. By this method, 98% of the 19,598 crash locations in Honolulu in 1990 have been successfully identified and the rest were identified manually.

Also, Austin (1995) developed a GIS-validation system for identifying the mis-located or mis-coded road crash records in the United Kingdom. The system used two databases, the first one contained the highway feature data (road centreline) and another one was the crash database. The location of each crash was plotted in GIS by using the five-figure grid references. The variables pertaining to each crash record were matched with the corresponding variables of the underlying highway features. The variables which were matched included road class, road number, district, speed limit, pedestrian crossing facilities, junction control, junction detail and carriageway type and markings. He termed all these "locational variables". A buffer zone of 24 m from either side of the centreline was created for each highway feature. When a crash fell within the buffer zone of a selected highway feature, the datum of the buffer zone was matched with