Detecting the spatial–temporal autocorrelation among crash frequencies in urban areas

Ali Akbar Matkan, Afshin Shariat Mohaymany, Matin Shahri, and Babak Mirbagheri

Abstract: Reducing the number and severity of crashes has been a primary concern of safety specialists. Identifying the crash patterns is an essential step in safety management since clustering of crash data helps the analysts to prioritize the safety improvements to such areas. The spatial essence of crash data particularly the presence of spatial autocorrelation reveals that crash occurrences are not only inclined to cluster in the same locations but within particular time intervals. This paper aims to detect spatial–temporal dependencies of crash occurrences using bivariate Moran’s I and local indicator of spatial association (LISA). Employing the yearly crash frequencies aggregated in 253 traffic analysis zone (TAZ) in Mashhad, Iran over 4 years (2006–2009), indicated that both bivariate Moran’s I and LISA yield significant patterns of clustering of crashes. The results of this analysis help the safety officials to sufficiently allocate the limited resources by prioritizing the detected clusters.

Key words: spatial–temporal autocorrelation, bivariate Moran’s I, LISA, urban safety.

Introduction

Intracity crashes linking with the growing number of vehicles has always been a concern of traffic engineers and safety specialists. Dealing with geographic data such as traffic crashes needs more attention since they display different properties than spatial data. The most important issue is the presence of spatial effects among neighboring samples over space or time which complicates the related analyses (Yamada and Rogerson 2003). The presence of spatial dependencies among geographic data such as traffic crashes often violates the assumption of independency that is implicit in most of the statistical models. Developing the spatial crash prediction models by taking the probable spatial dependencies into account.

Spatial autocorrelation measures the level to which extent the value of a variable at a specific location relates to the same value in proximate locations. When the level of interaction exceeds the expected level, the nearby locations have similar values and the autocorrelation is said to be positive. Conversely when the interaction is negative, the high values of variables are proximate with the low values and the spatial autocorrelation is negative. If data are located in space so that no relationship exists between the nearby values, the data are said to exhibit zero spatial autocorrelation. Univariate Moran’s I and local indicator of spatial association (LISA) are known as the most common methods to explore the existence of spatial autocorrelation among events and its adapted form known as bivariate Moran’s I and LISA explains the spatial pattern formed by two different variables (Anselin 1995; Anselin et al. 2007). Spatial–temporal autocorrelation is a special case in which the correlation of a variable in reference to spatial location of the variable within a time interval is assessed, i.e., correlation of a variable with itself over space and time (Anselin et al. 2002). While both univariate and bivariate Moran’s I aim to measure similarities and dissimilarities of spatial data, they are found to be less useful in the case of uneven spatial clustering.

Received 23 August 2012. Accepted 10 January 2013.

A.A. Matkan, M. Shahri, and B. Mirbagheri. GIS center, Faculty of earth sciences, Shahid Beheshti University, Tehran, Iran. A.S. Mohaymany. Faculty of civil engineering, Iran University of Science and Technology, Tehran, Iran. Corresponding author: Matin Shahri (e-mail: matinshahri@gmail.com).
Actually the global univariate and bivariate Moran's I might wrongly show that there is no relationship among the samples while there might be strong correlation in different parts of the study area hidden. Thus to explore extra information on the location of the clusters, univariate and bivariate LISA are used to detect local autocorrelation by determining exactly where the probable clusters exist in the study area (Anselin et al. 2002).

In the research followed by Khan et al. (2008) univariate Moran's I indicator was employed in GIS software to investigate whether neighboring counties in Wisconsin exhibited similar trends of ice-related crashes or not (Khan et al. 2008). Erdogan (2009) employed the spatial autocorrelation indicators Moran's I, Geary's C, and Getis-Ord General G to examine the dependency of the crash rates over the provinces in Turkey. The spatial distribution of provinces with high rates of crashes indicated nonrandom pattern and were detected as clustered with significance of p-value <0.05 based on spatial autocorrelation analyses (Erdogan 2009).

Investigating the spatial–temporal correlation among neighboring samples, e.g., dependency of total number of crashes or crash rates has been reviewed by different researchers as well. Wang and Abdel-Aty (2006) investigated the temporal and spatial correlation for longitudinal data and intersection clusters along corridors for the rear-end crashes at 476 signalized intersections in Florida counties and results of temporal analysis indicated the correlation of crash occurrences for years 2000 and 2002 (Wang and Abdel-Aty 2006). In the study conducted by Li et al. (2007) a GIS-based Bayesian approach was presented to investigate the spatial and temporal patterns of intracity motor vehicle crashes. The spatial analysis by day of the week and different years (1996–2000) suggested the stability in high-risk segments (Li et al. 2007). Erdogan et al. (2008) employed GIS methods to explore the crash hotspots in the city of Afyonkarahisar, Turkey to examine the hotspots conditions. The temporal analysis also showed the seasonal correlation among accidents particularly in crossroad of the villages, small cities and slippery areas. Also a higher number of crashes were detected in August and December. The daily analysis highlighted the correlation of crashes concentrating during the weekends, as well as on Fridays (Erdogan et al. 2008).

In another related study, spatial and temporal analysis of child pedestrian crash data was investigated by employing kernel density estimation in Santiago, Chile during the period 2000–2008. Subsequently, univariate Moran's I index test was employed to identify the probable spatial autocorrelation on crash occurrence over the time of day (Blazquez and Celis 2013).

In the recent paper presented by Eckley and Curtin (2012) the spatial–temporal clustering of fatal and injury incidents on a network was investigated using Knox method (Eckley and Curtin 2012). Although Knox index is a simple comparison of the relationship between incidents in terms of distance (space) and time, each individual pair of crash points is compared in terms of distance and time interval. Employing Knox such as other micro-level methods requires detailed individual data, the collection of which directly depends on the study area. In spatial analysis of crashes, several studies are found in which the crash behaviors have been investigated by developing micro-level models in different environmental settings, e.g., intersections, segments or corridors (Lee et al. 2003; de Smith et al. 2007; Anastasopoulos et al. 2008; Hildebrand et al. 2008; Cunto and Saccomanno 2009). The restriction of accessibility to large amounts of data needed for micro-level explanatory analyses, prompted the researchers towards developing macro-level models for which the data are aggregated over areal units, e.g., traffic analysis zones (TAZ) (de Guevara et al. 2004; Lovegrove and Sayed 2006), census blocks (Wier et al. 2009), county (Noland and Oh 2004) or ward (Noland and Qiddus 2004).

This study was conducted on Mashhad the second most populated city in Iran. Similar to other metropolises, the city is faced with the challenging problem of rapid socio-economic growth and the increasing number of vehicles during the last 2 decades. As a pilgrimage city, millions of people make a pilgrimage to Mashhad every year. The ever-increasing rates of trip generation in addition to high proportion of traffic in urban areas have increased the severity and number of crashes. Figure 1 illustrates the increasing trend of urban traffic crashes during the 4 years of study based on crash type classification.

In this paper we aim to explore the spatial–temporal autocorrelation of crash frequencies aggregated in 253 traffic analysis zones (TAZs) in Mashhad for four successive years from 2006 to 2009 and investigate whether the crashes demonstrate cluster or disperse patterns. The bivariate Moran's I and LISA, which had particularly been applied in the field of health (Uthman 2008; Hu and Ranga Rao 2009; Astutik et al. 2011) and ecology (Dray et al. 2008), will be employed for the first time in the field of safety. Combination of GIS and statistical analyses contributes to characterize the spatial effects of crashes and provides a quick view of the regions with the clustering of crash data that need more attention from transportation authorities. It also helps to appropriately decide on developing efficient crash prediction models.

![The increasing trend of crash occurrence](image-url)
Methodology

Data collection and process

This study is based on a database including the total number of crashes available for Mashhad during 4 years from 2006 to 2009, updated by Mashhad Transportation and Traffic Organization (MTTO 2007). According to the comprehensive transportation studies, the city has been divided into 253 homogenous TAZs. In the first step, the TAZ borders were investigated for any topological gaps or overlays and the geometrical errors were modified and eliminated. Crash data were then assigned to the TAZs using spatial join analysis and were grouped and allocated to the centroid of the TAZ. Assigning the crash data located along TAZ boundaries is the issue of concern. Since reporting the crashes in Iran is still in textual format, the geographical locations of crash points are not very exact. Furthermore, the TAZ borders and network centerlines are not precisely coincident (particularly by considering the highway widths). A query analysis in ArcGIS indicated that for each year of study, less than 0.1% of crashes corresponded to the TAZ borders for which we eliminated from analysis. In previous studies different suggestions were found. Fotheringham and Wenger (2000) pointed out that crash locations could be off by as much as 20 m by considering the street width, which could be large enough to overlap with one or more zones. Such inference is in line with the present study since the spatial autocorrelation measures the similarity of crash occurrences in neighboring TAZs whether they are assigned to a specific TAZ or its neighbors. In the research by Lovegrove and Sayed (2006), the crashes were single-counted and assigned to zones based on their geocoded locations. TAZ-level safety models were also developed in the studies of Hadayeghi et al. (2010), Naderan and Shahi (2010), and Abdel-Aty et al. (2011), but no discussion was done over the issue of assigning the crashes located along TAZ borders. As mentioned earlier, census blocks or regions can also be employed as spatial units; however since this research was considered as a pre-analysis to develop a spatial crash prediction model in which we aimed to employ travel demand as an underlying factor affecting the crash occurrence, therefore TAZ was selected. The spatial distribution of crashes over 253 TAZs in Mashhad within 4 years of study have been indicated in Fig. 2.

Global bivariate spatial autocorrelation

The global bivariate Moran’s I statistic quantifies the probable spatial dependency between two variables $x_i$ and $x_k$ in a same location $i$ (Anselin et al. 2002). This yields a counterpart of a univariate Moran-like spatial autocorrelation defined as follows:

$$ I_M = \frac{w^T z_i z_k}{n} $$

where $n$ denotes the number of observations and $w$ is the row-standardized spatial weight matrix that quantifies the relationships existing among neighboring samples. The weight matrix defines the neighbor set for each observation with nonzero elements for what is considered as neighbor and zero for others. Therefore since the spatial weights matrix imposes a structure on data, selecting a conceptualization that best reflects how features actually interact with each other is a crucial step in analysis. However, how to select weight function objectively is still a pending question remaining to be resolved (Chen 2009). Different spatial weight matrices such as inverse distance, inverse distance
squared, and contiguity polygons (rook or queen) are available $z_k$ and $z_t$ in eq. [1] and can be defined as

$$z_k = \frac{x_k - \bar{x}_k}{\sigma_k}$$

and

$$z_t = \frac{x_t - \bar{x}_t}{\sigma_t}$$

where $\bar{x}_k$ and $\bar{x}_t$ are the average value for a random distribution of variables and $\sigma_k$ and $\sigma_t$ are the corresponding standard deviations.

Most spatial statistical tests including Moran’s I begin by identifying a null hypothesis that essentially states that there is no spatial pattern among the features. In other words, the expected pattern (e.g., crash distribution) is just one of the many possible versions of complete spatial randomness. The $z$ score is a test of statistical significance that helps to decide whether or not to reject the null hypothesis. The related $p$-value also indicates the probability that the null hypothesis has falsely been rejected. Both $z$ score and $p$-value are associated with the normal distribution which relates to standard deviations with probabilities and allows significance and confidence to be inferred. The significance of bivariate spatial autocorrelation can be assessed commonly by means of a randomization (or permutation) approach. The randomization null hypothesis postulates that the observed spatial pattern of data represents one of many ($n!$) possible spatial arrangements. If it is possible to pick up the data values and throw them down onto the features in study area, one possible spatial arrangement will be achieved. The randomization null hypothesis states that if this exercise could be done (pick data up, throw them down) infinite times, most of the time we would produce a pattern that would not be markedly different from the observed pattern (our real data). Once in a while all the highest values might accidentally been thrown into the same corner of the study area, however the probabilities of doing that are small (de Smith et al. 2007). The resulting empirical reference will produce a Moran’s I scatterplot in which a spatial autocorrelation statistic as the slope of the regression line equal to the value obtained from eq. [1] is visualized. Moran’s scatterplot demonstrates the value of first variable on the vertical axis and the spatial lag (a weighted average of the value of second variable in the neighboring locations) on the horizontal axis (Anselin et al. 2002). Since the $z$ variables are standardized, the sum of squares used in the denominator of eq. [1] is constant and is equal to $n$. Therefore, the focus will be on the linear association between a variable $z_i$ at a location $i$ ($z_t$), and the corresponding spatial lag for another variable, ($w_{ij}z_j$) (this neighborhood is defined by spatial weight matrix). This concept centres on the extent to which values for one variable $z_k$ observed at a given location $i$ show a systematic (more than likely under spatial randomness) association with another variable $z_t$ observed at the neighboring locations (Wartenberg 1985).

**Space-time autocorrelation**

Space-time autocorrelation is a special case of mentioned bivariate spatial autocorrelation. Instead of employing different variables, $z_k$ and $z_t$ could be the same variable observed but in two instants of time, $t$ and $t’$. In this case, the bivariate Moran’s I computes the relationship between the spatial lag (defined by $w$), at time $t$ and the original variable, $z$, at time $t’$ or vice-versa. Therefore, this statistic quantifies the effect that a change in a spatial variable $z$, which operated in time $t’$ at an individual location $i$ ($z_t$), exerts over its neighborhood but at the time $t$. Hence, it is possible to define the global Moran’s I space-time autocorrelation statistic ($l_{i,t'}$) as follows:

$$l_{i,t'} = \frac{z_i w_{ij} z_j}{n}$$

where, as in the last case, the $z$ variables are also standardized. In the present study, $z_i$ will be the standardized crash frequency which is considered for the first year of study and $z_t$, the standardized crash frequency for the second year in neighboring TAZs. To define the spatial structure and type of neighborhood, different spatial weight functions such as inverse distance or contiguity weight matrix (rook or queen) can be defined. Although different weighting functions leads to producing the outliers that differ by defining one to another, the overall trend of expanding the clusters remains the same. To summarize, only the results obtained from defining the spatial weight matrix using inverse distance are demonstrated in the results section.

**Local Moran’s I and LISA**

The global Moran’s I will not provide the analysts with any information on where the clusters exist. As explained earlier it is likely that the global Moran’s indicator shows that there is no relationship among the data while the strong correlation in different parts of the study area exists. Local indicators of spatial association (LISA) provides a measure of association for each spatial unit (e.g., TAZ) and helps to identify the type and location of association (LISA) provides a measure of association for each spatial unit (e.g., TAZ) and helps to identify the type and location of spatial correlation (Anselin 1995). The numerator in eq. [4] can be decomposed into the contributions of the individual spatial units so that bivariate LISA is written as follows (Anselin et al. 2002):

$$l_{ij} = z_i' z_j$$

where $l_{ij}$ is the bivariate LISA, $z_i'$ implies the standardized value of the considered variable for the first time of study at location $i$, and $z_j'$ is the standardized value of the variable for the second time of study at locations $j$ (which are the neighbors of $i$ and their influence over $i$ is assessed by defining the spatial weight matrix).

In the present study, this statistic gives an indication of the degree of linear association (positive or negative) between the crash frequencies at ith TAZ in the first year of study and the weighted average crash frequencies at neighboring TAZs known as spatial lag in the second year of study. Greater similarity than indicated under spatial randomness suggests a spatially similar cluster in the two variables. However in the case of spatial-temporal analysis, the first and second variables are the attributes under study at specific locations but in different times. The results of such spatial-temporal analysis is a map that helps

**Table 1. Global bivariate Moran’s I for crash frequencies in years 2006, 2007, 2006, and 2009.**

<table>
<thead>
<tr>
<th>Original variable</th>
<th>Spatial lag</th>
<th>Bivariate Moran’s I</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>TAZ-level crashes in 2006</td>
<td>TAZ-level crashes in 2007</td>
<td>0.2348</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>TAZ-level crashes in 2006</td>
<td>TAZ-level crashes in 2008</td>
<td>0.1815</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>TAZ-level crashes in 2006</td>
<td>TAZ-level crashes in 2009</td>
<td>0.1422</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>TAZ-level crashes in 2007</td>
<td>TAZ-level crashes in 2008</td>
<td>0.2080</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>TAZ-level crashes in 2007</td>
<td>TAZ-level crashes in 2009</td>
<td>0.1493</td>
<td>&lt;0.05</td>
</tr>
<tr>
<td>TAZ-level crashes in 2008</td>
<td>TAZ-level crashes in 2009</td>
<td>0.1163</td>
<td>&lt;0.05</td>
</tr>
</tbody>
</table>
to identify the nature of autocorrelation over space and time so that the maps can be categorized into four groups: two categories of positive spatial autocorrelation, or spatial clusters (high–high and low–low) that relate to values physically surrounded by neighboring TAZs with similar values and two categories of negative spatial autocorrelation, or spatial outliers (high–low and low–high) that relate to values whose neighbor TAZs are dissimilar. Inferring the significant maps can be based on a permutation or randomization approach as explained earlier. The spatial analyses in this paper have been implemented through the GeoDa software (Anselin 2005) that provides a very user-friendly environment to implement Spatial Data Analysis (SDA) methods and is freely downloadable.

Results and discussion

Results of global bivariate Moran's I

As mentioned earlier, the objective of employing global bivariate Moran's I analysis is to indicate the variations of spatial distribution of data and in cases where the correlation of one variable is investigated for different time intervals, this results in the spatial-temporal autocorrelation. To this end, the bivariate Moran's I indicator was examined for a number of crashes that occurred in every TAZ in Mashhad for 1 year of study as the original variable and the spatial lag of the crash frequencies for the next year as the second variable; therefore the results implied the spatial-temporal autocorrelation. As indicated in Table 1, based on the analysis, T1
on the results of global bivariate Moran’s I, positive and significant (p-values <0.05) spatial autocorrelation across all spatial weights can be inferred. In simple terms, crash frequencies in Mashhad have some kind of organized spatial–temporal clustering patterns. Such systematic variation of crash occurrences would suggest that spatial–temporal effects should be taken into account for safety analyses. Figure 3 also depicts the corresponding global Moran’s I scatterplot. The slopes of the regression line in the six scatterplots is equal to Moran’s I indicator in Table 1, differentiate from zero which is an indicator for significant spatial–temporal autocorrelation among crash frequencies. The results obtained through implementing the bivariate Moran’s I indicator is consistent with Tobler’s first law of geography that states that geographic features that are near each other are likely to be more similar than distant features (Tobler 1970). It does not matter whether the influence is direct (e.g., the assessed value of a house being partly dependent on the value of surrounding houses) or indirect (adjacent TAZs having high number of crashes because crash occurrences are influenced by the set of spatially dependent variables such as population, network parameter, socio-economic...).

It is also worth noting that the analysis can be repeated if the values on the vertical and horizontal axes in the scatterplot are inversed, i.e., in Fig. 3a the crash frequencies in year 2007 as the original variable on the vertical axis and spatial weighted in year 2006 on the horizontal axis are introduced. For global statistics the results will be approximately the same.

Results of LISA

So far, the global analyses have suggested non-randomness in the overall spatial–temporal pattern of crashes. More information

Fig. 4. LISA map for crashes in comparing years (a) 2006 and 2007, (b) 2006 and 2008, (c) 2007 and 2009, (d), (e), and (f) the corresponding significant maps.
on what kinds of clustering may be present is provided by an analysis of LISA. The bivariate LISA cluster map for crash frequencies in year 2006 as the original variable and crash frequencies in years 2007, 2008, and 2009 as spatial lag have been indicated in Fig. 4. It is evident that the resulting bivariate LISA for most of the TAZs are highly statistically significant (p-value <0.05). The same analysis by inferring the bivariate LISA and the related significant maps for crash frequencies in year 2007 and 2008 as the original variable and crash frequencies in year 2008 and 2009 as the spatial lag has also been illustrated in Fig. 5. As evident the figure indicates the local patterns of clustering and spatial correlation between crash frequencies per TAZ in one year and the average number of crashes in the second year for its neighbor TAZs. The results of LISA inferred by local bivariate Moran's I enable dividing the study area into four sub regions corresponding to the clusters and outliers. These four categories match the four quadrants in the Moran's I scatterplot as shown in Fig. 3. The high–high and low–low regions (positive local spatial autocorrelation) represent spatial clusters, while the high–low and low–high locations (negative local spatial autocorrelation) represent spatial outliers. The high–high sub region is related to the areas having high (low) number of crashes in the first year of study which surrounded by the areas with high (low) weighted average of crashes for the second year of study (e.g., number of crashes in year 2006 and weighted average of number of crashes in year 2007 in Fig. 4a). It can be inferred from maps that the high–high sub regions corre-

Fig. 5. LISA map for crashes in comparing years (a) 2007 and 2008, (b) 2007 and 2009, (c) 2008 and 2009, (d), (e), (f) the corresponding significant maps.
sponding to the clustering of crashes have been concentrated mainly in the centre of Mashhad expanding from northwest to southeast. The stability of overall spatial pattern indicates the probability of crashes occurring in the same regions and that the crash frequencies follow organized patterns in Mashhad. The significant maps also reveal that for most of the TAZs located in high-high sub region, the results are highly significant (p-value <0.05). Another interesting result is that a majority of the units of measures that fall into the ‘‘Not-Significant’’ category, remains the same for all pairs of comparisons.

Since the results indicate the clustering of crash data, the areas that needed to be particularly targeted with safety-attention programs are explored. It can provide some guidance for decision makers and intervention planners where they should implement intervention action plans. Thus, the planners can appropriately allocate the limited resources (budget and time) for safety enhancements. Furthermore, it would be beneficial in selecting the appropriate crash prediction models particularly when the significant spatial dependencies prevail. The importance of the issue is yet more in case of developing crash prediction models, since when the spatial autocorrelation prevails, the results obtained by conventional nonspatial models might be misleading. This is in line with the findings of Levine et al. (1995), Miaou and Lord (2003), Agüero-Valverde and Jovaisan (2006), Eksler and Lassarre (2008), Quddus (2008), El-Basyouny and Sayed (2009), and Hadayeghi et al. (2010).

A point that must be emphasized is that what is detected as an outlier here may be correct or may be the result of some form of error (measurement, coding, representation etc.). Therefore it is essential to profoundly investigate what exactly is going on in a TAZ when it is detected as an outlier. Such locations could be of interest since they may represent the most important items in an investigation (e.g., the TAZ might include several intersections with increasingly higher crash frequencies) or might represent data that need to be removed or adjusted if either the information is known or suspected to be incorrect (de Smith et al. 2007). However, in case the original variable and the spatial lags are introduced conversely, LISA might encounter limitations but the results of detecting the clusters and overall behavior of events are not seriously influenced.

Conclusions

The key question investigated in this paper was whether the crash occurrences follow an organized spatial and temporal pattern or not. The consistent answer obtained in this paper is that the crash occurrences over TAZs in Mashhad were spatially, temporally dependent during the four years of study (2006–2009). This was yield through employing the bivariate global Moran’s I and LISA which denoted that the actual patterns of crashes were strongly associated spatially and temporally.

The main objective of this research was to indicate the importance of considering the spatial effects as the first step towards developing safety models. However, any safety analysis depends directly on various numbers of spatial variables including the socio-economic, exposure, land use, network parameters and travel demand which might vary over the space in study area. Developing a crash prediction model by exploring the varying effects of mentioned variables upon crash occurrence can also be fully investigated through some local methods such as geographically weighted regression techniques and will be topic of another research.

The data availability and the process of data collection is commonly a time-consuming procedure with some limitations towards the study area. By indicating the stable average behavior of urban crashes it is expected that if the safety prediction model is developed for a time with available data; the results can sufficiently be generalized to another time but with no or limited access to all detailed data needed for modeling.

Acknowledgment

Data for this paper was provided by Tarhe Haftom consulting engineers. Very special thanks and appreciation goes to Dr. Ali Naderan for his internuminally genuine collaboration in providing the data.

References

Anselin, L. 2005. GeoDa version 0.9.5.1.


