Periodica Polytechnica
Transportation Engineering

45(2), pp. 101-105, 2017
https://doi.org/10.3311/PPtr.9895

RESEARCH ARTICLE

Spatial Statistical Analysis of the Traffic Accidents

Tibor Sipos[1*]

Received 16 August 2016; accepted 18 August 2016

Abstract

26,000 people died on the roads of Europe in 2015, which is 5 500 less than in 2010. However, compared to the situation in 2014, no improvement could be experienced in this field at European Union level. Moreover, it was stated that 644 people died and 5575 suffered serious injuries as a result of road accidents on HU road network in 2015.

The objective of this study is to analyse the spatial clusters of traffic accidents on the secondary Hungarian road network in line with the No. 2008/96/EC Directive of the European Parliament and of the Council. Spatial statistical methods were used to analyse the statistical distribution of the traffic accidents. The investigated network was the sections of the secondary road network, while the investigated time horizon was between 2010-2012.

Keywords

road safety, infrastructure design, road infrastructure evaluation, spatial autocorrelation, accident analysis, hot spot analysis, black spot analysis, infrastructure safety

1 Introduction

The implementation of road safety projects with significant effect is a task with top priority due to the scarcity of available resources (Török, 2016). Thus there is reason for performing an effectiveness check on each project. During the effectiveness checks, besides the application of social-economic tests, a rather complex approach is necessary. Besides the financial analysis, the parameters affecting a wide range of economic operators from more aspects is also required to be involved and evaluated. Among others, such determining features are the frequency and social impact of transport accidents (Török, 2015).

A safety performance function by definition (SPF) is an equation to predict the number of fatalities per year at a given location as a function of roadway or intersection characteristics (e.g., number of lanes, traffic control, etc…). Mathematically, it is a model that can provide estimates about the events of accidents on certain road sections. Certain parameters of the road section and the traffic (e.g.: lane number, lane width, curve radius, traffic volume and traffic composition) are on the input side of the function. The expected frequency of accidents is on the output side. As a result of this, different SPFs can be assigned to different road segments. This assignment can be done by complex statistical software. Author constructed functions from historical data before the application of the given measures (Hauer, 1992). A statistical distribution that will provide the frequency of the events is required for the calculation of the SPF value. Typically, Poisson-type distributions are used for this purpose, with the assumption of having identical values and dispersions. However, it is typically shown by the observations that there is inconsistency between the Poisson distribution and the measurement data.

The Poisson regression – that is generally used for the establishment of a Safety Performance Function – is built on the identity of assuming an equality between the average and the variance of the dependent variable during our calculations. However, in reality, this assumption is frequently violated, consequently, the variance exceeds the average. This phenomena is called ‘overdispersion’ by the methodological literature (Kateri and Agresti, 2010). Usually there is one of the following two

[1] Department of Transport Technology and Economics,
Faculty of Transportation Engineering and Vehicle Engineering,
Budapest University of Technology and Economics,
H-1111 Budapest, Stoczek u. 2., Hungary

[*] Corresponding author, e-mail: tsipos@kgazd.bme.hu

reasons behind the overdispersion: (Yang, Hardin and Addy, 2010) On the one hand the explanatory variable that does not correlate with the independent variables in the model was excluded from the interpretation. On the other hand, there was interdependence between the observations. The concerns of crash frequency modelling include over-dispersion or under-dispersion of the count data, unobserved heterogeneity, spatial dependence, and the excess of zeros (Lord and Mannering, 2010; Mannering and Bhat, 2014)

Measured and modelled data have inconsistency, then modifications are required to the probability model. An overdispersion parameter is appropriate to use to decide whether the assumption of Poisson distribution is violated by the data series by which it can be expressed if the average value is exceeded by the distribution. By the use of the tool system for spatial statistics, the aim of this article is to justify our assumption according to which it can be stated that the application of a Safety Performance Function that uses the approach of Poisson-type distribution is limited on the secondary road network of Hungary.

2 Methodology

The results an appropriate framework were created based on the spatial autocorrelation for the analysis of the increasing statistical distribution of accidents. The investigation area has been limited to the secondary road network between 2010-2012. Author applied Varga's spatial autocorrelation definition (Varga 2002) on road safety patterns. The presence of spatial autocorrelation can be stated when: (1) the spatial grouping of quite similar values exists or, (2) the presence of quite different values by the neighbouring observation units exists. Our observations can be considered independent if neither of these criteria applies. Therefore, author have checked this criterium on road accident spatial dataset.

Local tests:

The possibility of local tests is provided by the Moran's tests by Anselin (Anselin, 1988), as well as by the Getis-Ord Gi* statistics (Getis, 2007).

The Getis-Ord Gi*:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \bar{x} \sum_{j=1}^{n} w_{i,j}}{s \sqrt{\frac{n \sum_{j=1}^{n} w_{i,j}^2 - \left(\sum_{j=1}^{n} w_{i,j} \right)^2}{n-1}}} \qquad (1)$$

where xj is the value of the jth element, $w_{i,j}$ is the spatial neighbouring matrix element between i and j, while the number of elements is signed by n.

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n} \qquad (2)$$

$$S = \sqrt{\frac{\sum_{j=1}^{n} x_j^2}{n} - \left(\bar{X} \right)^2} \qquad (3)$$

The $w_{i,j}$ matrix is a n X n symmetric matrix, with the jth element of its ith row expressing the strength of spatial relationship between the ith and jth spatial elements. The stronger the relationship, the higher the $w_{i,j}$ value is. By convention, a given point cannot be the neighbour of itself; there are 0 values in the main axis of the matrix, thus $w_{ii} = 0$ (Anselin, 1988).

It is emphasized by Getis (Getis, 2007) that filtering must be performed on each autocorrelated variable individually, depending on the optimal distance thereof (Bálint et al., 2014).

Consequently, the real cornerstone of spatial filtering is the determination of the optimal distance. The use of G (global) statistics was proposed by (Getis, 2007), searching for a distance where autocorrelation cannot be perceived yet. A similar geostatistical solution may be the application of semivariogram for each variable. (Friege et al., 2002) selected that particular distance parameter, by which the autocorrelation carried by the filtered variable was decreased to the minimum on the basis of the Moran's index. (Arbia et al., 2010) also relied on the beneficial features of Moran's I. They also regarded the distance with minimum significant Moran indicator appropriate; according to their theoretical considerations, this value is the threshold of the spatial spillover. (Bálint et al., 2014).

Global tests:

The index proposed by Moran in 1948 (Moran's index) shows whether the spatial distribution of currently examined data values refers to any regularity, that is, if the data of neighbouring unit areas are similar to each other (Izabella, 2011; Lafourcade and Mion, 2007). In case our data contain the location quotient or any other concentration index, we will receive the index of spatial autocorrelation between the concentration values.

$$I = \frac{n}{s_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} z_i z_j}{\sum_{i=1}^{n} z_i^2} \qquad (4)$$

where

z_i is the distance of x_i from the average of x, while z_j is the distance of xj from the average of x:

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} \qquad (5)$$

The value of the Moran's index may be in the [-1; 1] interval, its distribution is unknown, thus based on this value exclusively, the extent of autocorrelation in case of the spatial distribution of accidents cannot be determined. For in this case, with different levels of spatial distribution, different levels of spatial autocorrelation can be indicated by the I value. In the same way, the distribution of I values can also been influenced by the basic data. (Izabella, 2011). Thus, the distribution determined (estimated) by the Monte Carlo method is also required to be able to determine the spatial autocorrelation with the use of specific concentration values.

The GeoDa 0.9.5 or the ArcGIS 10.2 software programmes developed by Luc Anselin are both suitable for performing such calculations.

The z_I value for the statistics can be calculated in the following way:

$$z_I = \frac{I - E(I)}{\sqrt{V(I)}} \tag{6}$$

where

$$E(I) = -\frac{1}{n-1} \tag{7}$$

$$V(I) = E(I^2) - E(I)^2 \tag{8}$$

During the survey, all accidents were identified by one single point based on the planimetric coordinates of Descartes. In the present case these are the appropriate coordinates of the EOV HD72 Unified National Projection System.

3 Results

As the Moran's index is sensitive to the excess vacancy of the neighbour matrix, the net elements containing 0 element of accidents were separated.

The production of the standardized value of the Moran's I was determined by increasing the step rate of the d distance value of the neighbour matrix. The bottom limit of the d distance was 4 000 metres, which increased by 200-metre step rates up to 6 000 metres. The optimal distance when the effect of autocorrelation reaches its maximum value can be determined by this method for Getis' filtering. This prelim results were previously presented by the author.

Table 1 Determination of the global Moran's index with the examined distance values

Source: own edition

Distance threshold [m]	Moran's Index:	z-value	p-value	Variance
4 000	0.243905	52.988542	>0.01	0.000021
4 600	0.212650	53.434475	>0.01	0.000016
5 000	0.204887	55.099081	>0.01	0.000014
6 000	0.176693	54.800135	>0.01	0.000010

The maximum z value determined by the 200-m step rating was measured at a distance of 5 000 metres. Subsequently, the maximum z value was measured at 5 030 metres by fine-tuning. 55.099093 z-score and 0.204887 Moran's Index belonged to the d distance of 5 030 metres.

As the value of z became 55.099, it can be said that the chance of being randomly clustered is less than 1%.

Consequently, it can be clearly concluded that the spatial elements formulated by the spatial aggregation of accidents that occurred on the secondary road network between 2010 and 2012 are autocorrelated, thus statistical analyses can be performed in view of this exclusively.

A local Getis-Ord G* statistical analysis can be performed subsequently to the determination of the maximum d distance, thus the significance of specific local groupings (clusters) can be determined. The Gi z and Gi p values of the fishnet element (area of 1 km2) related to each section of the secondary road network were determined by the analysis.

The so-called Hot and Cold Spots can be clearly distinguished based on the z and p values.

Based on these, author has distinguished 6 classes that are respectively: Cold Spot with a significance level of 99%, Cold Spot with a significance level of 95%, Cold Spot with a significance level of 90%, Hot Spot with a significance level of 99%, Hot Spot with a significance level of 95%, Hot Spot with a significance level of 90%, not significant.

The strength of spatial relationships is substantially identified by the Gi statistics with the concentration of weighted spatial points. The value of the Gi will be high if the values that are higher within a given distance are clustered, while it will be low if low values are concentrated (Bálint, 2011). The difference between the observed and expected values of the Gi statistics can provide a reply to the question of whether the clusterization of the high or low value of the variable is characteristic of the environment of the given location (Bálint, 2011). With an appropriate consideration and the use of new theories, it can be stated about each specific location whether that section can be called a 'hot' point from the aspect of accidents or not. Subsequently, the weight matrix with the principle of K-nearest neighbours (KNN) was applied instead of the spatial locations of accidents and the distance-based weight matrices. The loss values of accidents that can be related to certain parts of the sections were determined (Holló and Hermann, 2013). Then a point network was developed by the spatial join of the ratio of the loss value and the traffic volume to the centroids of the subsections. The possible combinations of clusters were determined by LISA's parameters (LISA – Local Indicators of Spatial Association), subsequent to the production of the new weight matrix and the determination of the global G statistics – Global Moran's I. The GeoDa 0.9.5 software developed by Luc Anselin was used for the analyses. The statistical significance level was determined using a complex Monte Carlo randomization procedure. The loss value is assessed by comparing the actual value to the value calculated for the same location by randomly reassigning the data among all the areal units and recalculating the values each time. If an actual LISA score is among the top 5% of scores associated with that location under randomization, then it is judged statistically significant at the 0.05 level.

Actual data values across space are compared to data values that are randomly generated and randomly spatially distributed by a Monte Carlo procedure. Ultimately, the LISA stat indicates local clustering or local outlier areas.

For analytical purposes the high-high cluster can be regarded as the most significant, thus those subsection centroids where the ratio of accidents that have high loss value from accidents compared to the traffic are significantly concentrated.

It was determined by the spatial autocorrelation index of the global Moran's I that regularity can be observed in the distribution of road accidents on the secondary road network of Hungary between 2010 and 2012: related to the frequency of accidents, there is positive spatial autocorrelation in the whole period under survey. The accident rate of sections located in the vicinity of sections with high frequency of accidents is similarly high, while the ones located in an environment with lower accident rate have low accident frequency.

4 Conclusion

Based on the presented results, we can clearly determine a spatial autocorrelation between road accidents, which justifies our hypothesis that the application of a Safety Performance Function using the approach of Poisson-type distribution is limited on the secondary road system of Hungary. Thus, a statistical analysis of the data on accidents is proposed by the author before the preparation of the SPF. In case the data can be characterized by spatial interdependence, this is likely to be the reason behind the overdispersion that appeared in the applied modelling procedure.

This is because, one of the assumptions of the Poisson distribution is that the observations are independent from each other, thus one accident does not have an effect on the probability of developing another one. If this condition is not fulfilled, based on the Poisson distribution, the frequency of high and low-value elements will be higher than it was expected. In consequence of this, the dispersion of the variable will rise. The most serious consequence thereof is that the estimated standard errors of the coefficients will be less than the actual ones and due to this underestimation, the picture suggested by the significance tests will be false, it will be much more favourable than the reality (Moksony, 2006).Thus, we can easily receive a result, which is statistically significant, in spite of the fact that our assumptions are actually incorrect. During the determination of standard errors, we consider the average and the variance as identical, which may lead to a false result by our SPF function. The econometric modelling of autocorrelation can be a way of eliminating the problem of overdispersion (by the model of spatial delay or the autocorrelation model of a spatial error), while another type of possibility can be the option of non-parametric spatial filtering.

Transportation projects have a range of such individual properties, which distinguish them from the investments in other fields of the economic life. The priority of the professional preparation of decisions is especially emphasized by these properties as such projects are typically implemented at high investment costs, the lifespan of these projects can be more decades and an investment that was inappropriately prepared for can be modified with the application of a high amount of resources in the future. The determination of social-economic external cost values and the frequency of accidents is one of the necessary conditions of evaluating projects. In the international practice, the use of an appropriate SPF is common. The methodology introduced by the author provides suitable framework for selecting an SPF that is appropriate to the particular distribution by the examination of the spatial autocorrelation of accidents. Based on the examination of the events of accidents on the secondary road system of Hungary, it can be stated that these are clustered and have a type of spatial 'density', thus, the distribution that will be applied during preparation of an SPF can be selected in accordance with this.

References

Anselin, L. (1988). A test for spatial autocorrelation in seemingly unrelated regressions. *Economics Letters*. 28(4), pp. 335-341. https://doi.org/10.1016/0165-1765(88)90009-2

Arbia, G., Battisti, M., Di Vaio, G. (2010). Institutions and geography: Empirical test of spatial growth models for European regions. *Economic Modelling*. 27(1), pp. 12-21. https://doi.org/10.1016/j.econmod.2009.07.004

Bálint, L. (2011). *A születéskor várható élettartam nemek szerinti térbeli különbségei*. (Spatial Gender Differences of Life Expectancy at Birth). KSH, Budapest. (in Hungarian)

Bálint, L., Daróczi, G., Bozsonyi, K., Tóth, G. (2014). A választói viselkedés térbeli modellje – empirikus kísérlet budapesti adatok alapján. (Spatial Model of Voting Behaviour –Empirical Study Based on Budapest Data). *Tér és Társadalom-Space and Society*. 28(3), pp. 32-49. (in Hungarian) https://doi.org/10.17649/TET.28.3.2591

Friege, L., Hoell, D., Ferstl, R., Leplow, B., Aldenhoff, J. (2002). Impaired spatial learning in schizophrenic patients. *European Psychiatry*. 17 (Suppl. 1), p. 158. https://doi.org/10.1016/S0924-9338(02)80684-8

Getis, A. (2007). Reflections on spatial autocorrelation. *Regional Science and Urban Economics*. 37(4), pp. 491-496. https://doi.org/10.1016/j.regsciurbeco.2007.04.005

Hauer., E. (1992). Empirical Bayes approach to the estimation of "unsafety": the multivariate regression method. *Accident Analysis & Prevention*. 24(5), pp. 457-477. https://doi.org/10.1016/0001-4575(92)90056-O

Holló, P., Hermann, I. (2013). A közúti közlekedési balesetek által okozott társadalmi-gazdasági veszteségek aktualizálása. (Actualization of Social-Economic Losses Caused by Road Accidents.) *Közlekedéstudományi Szemle*. 63(3), pp. 22–27. (in Hungarian)

Kateri, M., Agresti, A. (2010). A generalized regression model for a binary response. *Statistics & Probability Letters*. 80(2), pp. 89-95. https://doi.org/10.1016/j.spl.2009.09.016

Lafourcade, M., Mion, G. (2007). Concentration, agglomeration and the size of plants. *Regional Science and Urban Economics*. 37(1), pp. 46-68. https://doi.org/10.1016/j.regsciurbeco.2006.04.004

Lord, D., Mannering, F. (2010). The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*. 44(5), pp. 291-305. https://doi.org/10.1016/j.tra.2010.02.001

Mannering, F. L., Bhat, C. R. (2014). Analytic methods in accident research: Methodological frontier and future directions. *Analytic Methods in Accident Research*. 1, pp. 1-22. https://doi.org/10.1016/j.amar.2013.09.001

Moksony, F. (2006). A Poisson-regresszió alkalmazása a szociológiai és demográfiai kutatásban. *Demográfia*. 49(4), pp. 366-382. (in Hungarian)

Szakálné Kanó, I. (2011). A gazdasági aktivitás térbeli eloszlásának vizsgálati lehetőségei (Possible Surveys on the Spatial Distribution of Economic Activities). *Statisztikai Szemle*. 89(1), pp. 77-100. (in Hungarian)

Török, Á. (2015). Analysing the Connection of Hungarian Economy and Traffic Safety. *Periodica Polytechnica Transportation Engineering*. 43(2), pp. 106-110. https://doi.org/10.3311/PPtr.7953

Török, Á. (2016). Statistical Analysis of a Multi-Criteria Assessment of Intelligent Traffic Systems for the Improvement of Road Safety. *Journal of Finance and Economics*. 4(5), pp. 127-135. https://doi.org/10.12691/jfe-4-5-1

Yang, Z., Hardin, J. W., Addy, C. L. (2010). Score tests for overdispersion in zero-inflated Poisson mixed models. *Computational Statistics & Data Analysis*. 54(5), pp. 1234-1246. https://doi.org/10.1016/j.csda.2009.11.010