Modeling accident hotspots to locate roadside equipment based on intelligent transportation system

Saman Shafipour*, Mahmoud Reza Delavar **, Abbas Babazadeh ***

* GIS Department, School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran, Email: saman.shafipour1371@gmail.com

** Center of Excellent in Geomatic Eng., School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran, Email: mdelavar@ut.ac.ir,

*** School of Civil Eng., College of Engineering, University of Tehran, Tehran, Iran, Email: ababazadeh@ut.ac.ir

Abstract. Road transport has always attracted immense attention in Iran's planning as one of the main transportation system and economical infrastructures. High number of accidents and road fatalities in Iran reveals the weak safety in Iran's roads, therefore in the era which is called digital era, information technology is a necessary and efficient tool to help the management of transportation and increasing the road safety.
Geospatial information system and intelligent transportation system are among the branches of information technology that are used in transportation management. Intelligent transportation system is composed of various components with different applications can be used in all of the transportation systems. On the other hand installing and setting up the components of this system is very expensive, justifying the accurate and proper location

LBS 2021
16th Conference on Location Based Services
24-25 November 2021 Online

Published in "Proceedings of the 16th International Conference on Location Based Services (LBS 2021)", edited by Anahid Basiri, Georg Gartner and Haosheng Huang, LBS 2021, 24-25 November 2021, Glasgow, UK/online.

https://doi.org/10.34726/1757 | © Authors 2021. CC BY 4.0 License.

determination for these facilities.

To the best of our knowledge, Empirical Bayesian Kriging Regression Prediction (EBKRP) and Forest-based Classification and Regression to model the accident prone areas and predict the hotspots has not been undertaken so far. In addition, preparing, preprocessing and exploring the impact of input variables which varies in number and nature in ArcGIS Pro software has not been done.

Contribution of this research is in predicting high risk areas as an appropriate place for the installation of intelligent speed bumps using machine learning methods and data mining based on artificial intelligence in a locational intelligence field.

The data used in this study consist of the official data of the traffic accidents in the period 2018 to 2019 which are available in the accidents and road transport system and has been obtained using programming in the web environment, intelligent descriptive information obtained from smart cameras of the video surveillance in the context of locational information system, non-intelligent descriptive information obtained from the Ministry of Roads and Urban Development related to the characteristics of suburban roads of Mazandaran Province in the North of Iran and also TanDEM digital elevation model with a spatial resolution of 12.5 meters.

In order to predict high traffic accident risk areas, first the area of Mazandaran Province was divided into a number of hexagons to reduce the error effect of the data fusion process. The accidents data and the surrounding land uses and land covers have been extracted from the images acquired from the smart transportation monitoring cameras.

For predicting the dependent variable and estimating the coefficients of significance considering the available data uncertainty, an automatic method has been proposed in this research. The method is based on heuristic regression, ordinary least squares regression and spatial rhythmic regression by considering distance as an independent variable and regression and forest-based classification with a combination of raster, vector and artificial data were used as independent variables. In addition, a new method of predicting EBK regression with raster format was proposed and implemented in this paper. Heat mapping tools have been used to convert vector variables into raster format. The integration of DEM as a variable containing ground height information with other inputs of the EBKRP method was also employed.

Furthermore, the combination of digital elevation layer was used as a variable containing information about the earth with other inputs of the EBKRP method.

The results showed that the information variable obtained from smart images in the training regression process and forest-based classification methods are among the effective variables in the modeling and predicting high traffic accident risk areas.

In addition, it is shown that the residuals obtained from the spatial statistics employed methods have a random distribution. On the other hand, based on the validation performed for each of the implemented methods, it was found that the adjusted coefficient of determination (Adjusted-R2) for spatial rhythmic regression method has been increased compared to those of the normal least squares regression, regression method and forest-based classification. ۲۰% of the data were selected to validate the results and the mean square error (MSE) was estimated to be 0.012. The Geostatistics toolkit in the two cases has been used in terms of time. The cross-validation method employed showed that in the case of considering the digital layer of height in the modeling process, the accuracy of the model prediction process has been improved.

Keywords. Intelligent Transportation System, Geospatial Information System, EBK Regression Prediction

1. Introduction

The use of geospatial information system (GIS) in the transportation network has been developed in recent years, where GIS transport (GIS-T) is quite common in this field. Therefore, the practical principles of using geospatial information system science and technology in transportation related issues has been widely welcomed by the transportation community (Agyemang 2013).

On the other hand, intelligent transportation systems (ITS) use various technologies such as navigation systems, electronic toll payment systems, traffic control sensors, weather monitoring devices, electronic messaging boards. In order to improve the transportation systems and enhance the infrastructure of the road equipment, collecting, storing, analyzing and integrating the qualified transportation information is necessary. Furthermore, the information in the transportation system, such as road networks are spatially referenced (Khan, Rahman et al. 2017, Zhu, Yu et al. 2018).

2. LITERATURE REVIEW

Previous research on road accidents has focused on spatial factors affecting the distribution and type of accidents in urban and suburban roads(Shafabakhsh, Famili et al. 2017). Accumulation of accidents in any place can not only be affected by human factors and vehicle defects, but also the characteristics of the road, usages of their surrounding facilities, the frequency and type of accidents and the amount of the associated damages (Shafabakhsh, Famili et al. 2017, Yuan, Zhou et al. 2018).

The analysis of these effects can be used in future road safety planning, including determining the appropriate location for the construction of intelligent accelerators or the proper distribution of conventional accelerators in the absence of proper road intelligent infrastructure, with the aim of reducing accidents by road transport authorities.

3. RESEARCH GAPS AND QUESTIONS

The data may be flawed for a variety of reasons. For example, data may be lost in some cases because a sensor may be temporarily disabled, a sampling point may be inaccessible, or data values may be deliberately hidden for confidential protection. When one or more values are missing, most statistical methods in data preprocessing remove that attribute from the process by default. To bridge this gap, which implies the presence of areas with no data, by filling in the missing values, the nulls are estimated with values based on spatio-temporal data analysis with respect to their neighborhood to minimize the effect of those values with missing data(https://pro.arcgis.com/en/pro-app/latest/tool-reference/space-time-pattern-mining/fillmissingvalues.htm).

Data diversity is another issue in the intelligent transportation system, which is collected in various formats and in different ways including numerical data obtained from sensors on vehicles and roads and textual data obtained from the media. Social and image data contained in the maps are collected in the geospatial information system. This data can be organized from semi-structured data (e.g., text reports, images, videos, and audio files) to structured data (e.g., data received from smart devices such as video surveillance cameras, sensors and traffic accident data in a database) (Khan, Rahman et al. 2017).

Therefore, data infrastructures and systems that can model and predict large volumes of data are needed to convert the equipment information from a technology-based system to a complex data-driven system, such as a geospatial information system (Khan, Rahman et al. 2017).

4. Methodology

Machine learning is a branch of artificial intelligence in which structured data is processed by an algorithm to solve a problem. Deep learning is a subset of machine learning that uses multiple layers of algorithms in the form of neural networks. Input data is analyzed through different network layers, each layer defining specific features and patterns in the data(L. Bennett).

EBKRP is a geostatistical interpolation method employed in ArcGIS Pro software that combines the Empirical Bayesian Kriging (EBK) method with regression analysis. In the regression models, heuristic variables are often related to each other. The problem of multilinearity is solved by converting the primary independent variables necessary in the raster model to their main components before constructing a regression model (Malcheva, Bocheva et al. 2019).

Official and raw data related to suburban accidents in Mazandaran Province, for the years 2018 and 2019 from the comprehensive information system of accidents and transport accidents in Iran, which has been recorded by the experts of the Roads and Transportation Organization, in the form of a set of Excel files. Accident data for two years were obtained with the JavaScript programming language. After the preprocessing steps, the ground was referenced by the linear reference system method. Because the recorded information from accidents may have been incomplete and subject to change, an indicator is used to identify accident hotspots with fatalities with those accidents that only cause damage which are considered for the road traffic safety analysis.

The distribution and dispersion of land use density, especially educational land uses along the roads in Mazandaran Province, have a significant impact on traffic accidents. Therefore, various information including text reports, roadside user information (e.g. educational and medical land uses) as well as digital model of land height and spatial features (e.g. roads and population centers) are collected and aggregated in a spatial database.

Optimal performance were determined by ordinary least squares (OLS) and geographically weighed regression (GWR) methods with R^2 and AdjR^2 values, Forest-based Classification and Regression methods with R^2 and MSE values, as well as EBK Regression Prediction (EBKRP) and EBKRP + digital elevation model (DEM) methods with Continuous Ranked Probability Score (CRPS) and Root Mean Square Standardized Error (RMSSE) values.

5. Discussion

In order to achieve one of the objectives of the research based on the use of GIS-based emerging machine learning methods, it was determined that from the set of spatial statistics tools, a prediction model based on Forest-based Classification and Regression due to the diversity in accepting the number and nature of input information, as well as from the Geostatistics toolkit, the EBKRP-based prediction model due to the acceptance of terrestrial raster data such as DEM are suitable models for locating the intelligent accelerators. Figure (1) is a graphical output of areas appropriate to the installation of intelligent accelerators or areas predicted by the EBKRP method.

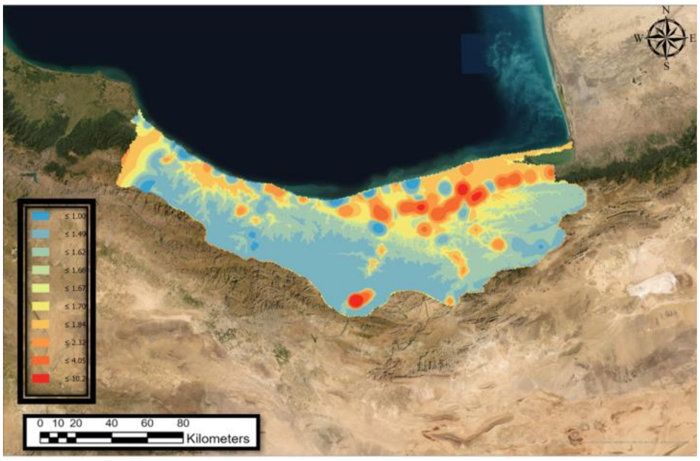

Figure 1. The spatial distribution of some car accidents traffic hot spots

The disadvantages of the exploratory regression and the least squares methods are the limitation in the number of input variables. To best of my knowledge, the limitations of the machine learning methods in GIS can be mentioned as the number as well as the unacceptance of the input variable with raster format (Ali, Sajjad et al. 2021). Therefore, the EBKRP method was selected to determine the importance of variables and the prediction was made based on 25805 hexagonal tessellations employed in dividing the Mazandaran Province. The accuracy of the proposed method based on the coefficient of determination was 0.915 and the error equal to 0.012 was obtained.

6. Conclusion

In comparison with(Effati, Rajabi et al. 2015) who used the classification tree method and fuzzy regression to predict the model, in the present study the forest-based classification and regression method was used. The advantage of this method is in the final predictions that are not based on any single tree but on the whole forest.

References

Agyemang, E. (2013). "A cost-effective Geographic Information Systems for Transportation (GIS-T) application for traffic congestion analyses in the Developing World." Ghana Journal of Geography **5**: 51-72.

Ali, G., M. Sajjad, S. Kanwal, T. Xiao, S. Khalid, F. Shoaib and H. N. Gul (2021). "Spatial–temporal characterization of rainfall in Pakistan during the past half-century (1961–2020)." Scientific reports **11**(1): 1-15.

Effati, M., M. A. Rajabi, F. Hakimpour and S. Shabani (2015). "Prediction of crash severity on two-lane, two-way roads based on fuzzy classification and regression tree using geospatial analysis." Journal of Computing in Civil Engineering **29**(6): 04014099.

https://pro.arcgis.com/en/pro-app/latest/tool-reference/space-time-pattern-mining/fillmissingvalues.htm.

Khan, S. M., M. Rahman, A. Apon and M. Chowdhury (2017). Characteristics of intelligent transportation systems and its relationship with data analytics. Data analytics for intelligent transportation systems, Elsevier**:** 1-29.

L. Bennett, M. l. i. A., " ed: ArcUser, 2018.

Malcheva, K., L. Bocheva and T. Marinova (2019). "Mapping temperature and precipitation climate normals over Bulgaria by using ArcGIS Pro 2.4."

Shafabakhsh, G. A., A. Famili and M. S. Bahadori (2017). "GIS-based spatial analysis of urban traffic accidents: Case study in Mashhad, Iran." Journal of traffic and transportation engineering (English edition) **4**(3): 290-299.

Yuan, Z., X. Zhou and T. Yang (2018). Hetero-convlstm: A deep learning approach to traffic accident prediction on heterogeneous spatio-temporal data. Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.

Zhu, L., F. R. Yu, Y. Wang, B. Ning and T. Tang (2018). "Big data analytics in intelligent transportation systems: A survey." IEEE Transactions on Intelligent Transportation Systems **20**(1): 383-398.