Physica A 520 (2019) 309–316

Contents lists available at ScienceDirect

Physica A

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

Transportation hazard spatial analysis using crowd-sourced social network data

Ali J. Ghandour [a,*], Huda Hammoud [b], Luciano Telesca [c]

[a] National Council for Scientific Research (CNRS), Beirut, Lebanon
[b] American University of Beirut, Beirut, Lebanon
[c] Institute of Methodologies for Environmental Analysis, National Research Council, Tito Scalo, PZ, Italy

HIGHLIGHTS

- Analysis of accidents' types distribution is provided.
- Spatial autocorrelation in the Lebanese accidents dataset and high clustering accidents areas is detected.
- Hot spots variation between the summer and winter seasons is inspected.
- Road hazard index is proposed to measure road segments risk analysis.

ARTICLE INFO

Article history:
Received 28 May 2018
Received in revised form 14 August 2018
Available online 21 January 2019

Keywords:
Spatial analysis
Crowd-sourcing
Car crash map
Hazard assessment

ABSTRACT

The safety hazard and the additional costs on transportation due to road accidents invite the necessity to minimize their impact. In this paper, we study the spatial-clustering behavior and hazard vulnerability of car accidents that occurred in Lebanon between 2015 and 2018. Assessment of spatial clustering of accidents and hot spots densities were examined using the Global G method of spatial autocorrelation and Getis–Ord G_i^* statistics. A novel Road Hazard Index (H_i) was proposed to assess hazard vulnerability of road networks and to develop a road hazard prediction model.

1. Introduction

Developing countries are riddled with hazardous roads so that the probability of dying from a vehicle crash is greater than from most natural causes of death [1,2]. Lethal road accidents are attributed to impairment in the road transport system. Lebanon, not being an exception, lacks a sustainable transport system and infrastructure, provoking the level of accidents on the roads. The high portion of traffic accidents highlights the necessity of a comprehensive study of road's safety to identify unsafe roads and present hazard model. Groundwork of the solution is to assess the road network and identify the malfunction roads where a high frequency of accidents occur then target these roads for safety upgrading with affordable engineering remedies.

Given the poor quality of the reported accident statistics base and the lack of statutory authority in charge of collecting and reporting accidents' data, one used to resort to anecdotes and specific case studies for clues. To refine accident data collection, we previously proposed a Lebanese Road Accident Platform (LRAP) [3] which is a real-time online platform that collects crash events from social media following a crowd-sourcing model.

* Corresponding author.
E-mail addresses: aghandour@cnrs.edu.lb (A.J. Ghandour), hah57@mail.aub.edu (H. Hammoud), luciano.telesca@imaa.cnr.it (L. Telesca).

https://doi.org/10.1016/j.physa.2019.01.025
0378-4371/

Table 1
Distribution of road traffic accidents based on crash type.

Crash type	Vehicle–vehicle	Vehicle–pedestrian	Vehicle–motorcycle
Count	4916	1876	1504
Percentage	55.13%	21.04%	16.86%

This paper focuses on the use of the accidents data extracted from the Lebanese Road Accident Platform (LRAP) to study the spatial-clustering behavior and trends of car crashes and develop an accident hazard vulnerability index of the roads segments in Lebanon.

The remainder of this paper is organized as follows: Section 2 discusses the collected data and provides initial analysis based on defined crash types. In Section 3, spatial analysis is applied to the crash events by testing the spatial autocorrelation and applying hot spot analysis. Section 4 introduces Hazard vulnerability analysis which is adapted by defining a road hazard index to road segments and inferring the hazard index for roads with uncollected data. Finally, Section 5 presents concluding remarks.

2. Car accident database

We obtained the data presented in this study from the publicly available database we built in the scope of a national project (http://sctl.cnrs.edu.lb:8000/pyfyp/plot_twitter), that aims to produce a geographical database of accidents in Lebanon [3]. The platform produces a geographical database of the accidents occurring in Lebanon by crowdsourcing reported accident statistics on social media from three credible governmental sources which are the Traffic Management Center, Civil Defense and Lebanese Red Cross. The database contains information about the accidents such as the location, time, type of accident, the number of injuries, and the number of lethal victims.

Fig. 1 shows the spatial map of the entire Lebanese database for car accidents covering the period of three years from February 2015 February 2018. The majority of vehicle crashes occurred in Beirut and Mount Lebanon. Beirut is the capital and the center of finance and trade in Lebanon that accommodates most of the governmental institutions and major private sector companies. Out of the 4.5 million Lebanese inhabitants, 1.3 million live in the Greater Beirut Area, causing the capital to be densely populated [4]. Another a half million Lebanese students and workers commute daily to the capital [5]. The burdensome traffic, precarious road conditions and erratic driving explain the concentration of vehicle accidents in Beirut and Mount Lebanon, as evident in Fig. 1. Moreover, Fig. 2 shows a zoomed view of Beirut city from the studied dataset where Fig. 2a displays the spatial location of crash events and Fig. 2b shows the daily number of crash events shown in Fig. 2a.

To analyze the various types of car crashes taking place at Lebanese roads, we defined the following five categories of crash types: (*i*) Vehicle–Vehicle, (*ii*) Vehicle–Pedestrian, (*iii*) Vehicle–Motorcycle, (*iv*) Vehicle–Infrastructure and (*v*) Others.

We computed the frequency of accidents across these different crash types categories to find that more than half of the occurring accidents are vehicle crashing against another vehicle, followed by vehicle hitting a pedestrian passing through (21.04%) and vehicle crashing with a motorcycle (16.86%) as shown in Table 1.

A closer look at the Vehicle–Pedestrian crash type reveals that the most of occurrences were in Saida (28%), (21%), Moussaytbeh (19%), Tyre (18%) and Zahlé (14%). Urban environments are often black spots for pedestrian crashes given the highly condensed and heavily trafficked areas contributed by the urban design [6]. Risky behavior exhibited by the pedestrians such as ignoring traffic signals and sidewalks reflects cultural attitudes which promote risk taking behavior with long-term impact on vehicle–pedestrian crash frequencies [7].

Motorcycle riders represent the most vulnerable road users in Lebanon, especially with the common lack of minimal safety conditions such as motorcycle helmet. The accidents involving motorcycle crashes are the third top crash type in Lebanon, where Mousseytbeh ranks with the largest number of motorcycle-involved accidents (29%) followed by Tyre (27%), Tripoli (16%), Ras Beirut (15%) and Chiah (13%).

We can attribute the large number of motorcycle-involved accidents to the overcrowded and poor living circumstances manifested by the lack of decent job opportunities and low Gross Domestic Product (GDP) of those areas. Mousseytbeh, for instance, contains Wata camp which is one of the poorest slums in Beirut. In Wata camp, house renting is cheap and occurs mostly through illegal processes for very cheap cost (around 6.5$/month) [8]. Moreover, Tyre contains two large refugee camps which are Rashidieh and Al-Buss camps, and for similar reasons as Mousseytbeh, one can observe many motorcycle accidents occurrences.

3. Spatial analysis

3.1. Spatial autocorrelation: Global G statistic

To inspect the spatial dependencies among the crash frequencies, we employ the Getis–Ord General *G* statistic as an indicator of spatial autocorrelation [9].

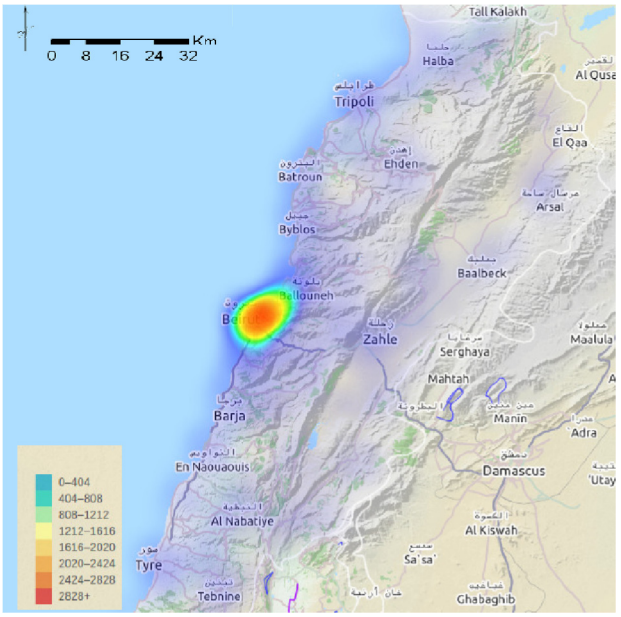
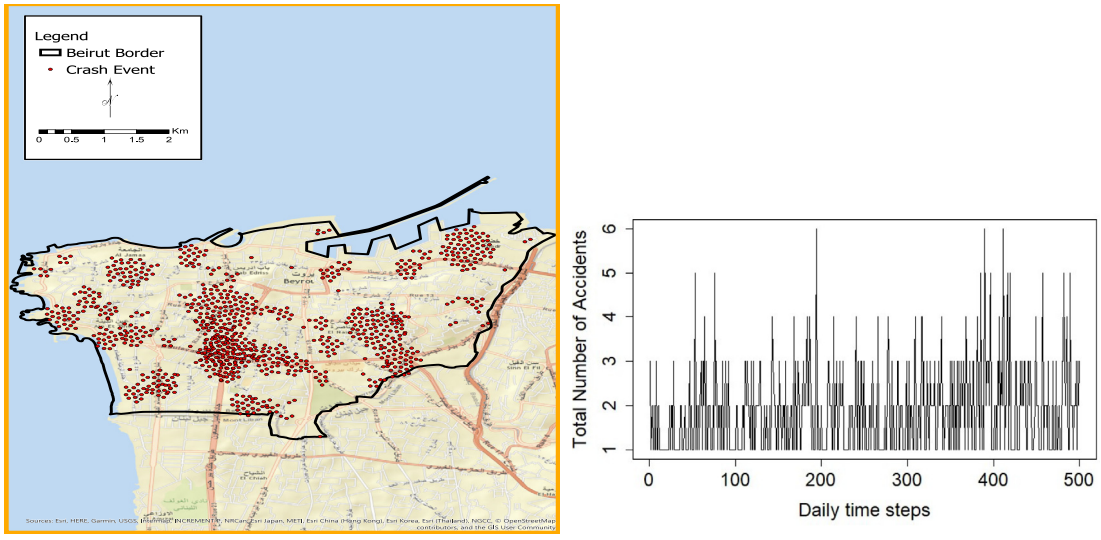


Fig. 1. Spatial distribution of Lebanese car accidents from 2015 until 2018 [33.8547° N, 35.8623° E].



(a) (b)

Fig. 2. Figure 2a shows the location of crash events in Beirut and Figure 2b shows the time occurrence of those crash events.

The General G statistic of the overall spatial association is given as follows:

$$G = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_i x_j}{\sum_{i=1}^{n}\sum_{j=1}^{n} x_i x_j}, \forall j \neq i, \tag{1}$$

where x_i and x_j are attribute values and w_{ij} corresponds to the spatial weight values. n is the number of features in the dataset and $\forall$ j $\neq$ i indicates that features i and j cannot be the same feature.

The normalization of G, z_G-score is given as:

$$Z_G = \frac{G - E[G]}{\sqrt{V[G]}}, \tag{2}$$

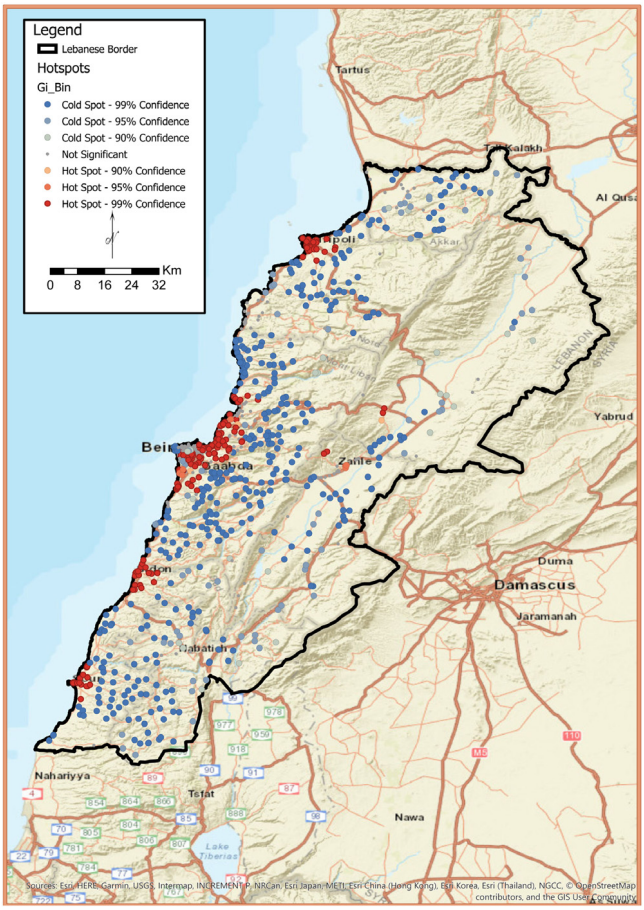


Fig. 3. Hot Spot analysis of Lebanese crashes occurrence frequency [33.8547° N, 35.8623° E].

where $E[G]$ and $V[G]$ are defined in Eqs. (3) and (4), respectively:

$$E[G] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}}{n(n-1)}, \forall j \neq i \tag{3}$$

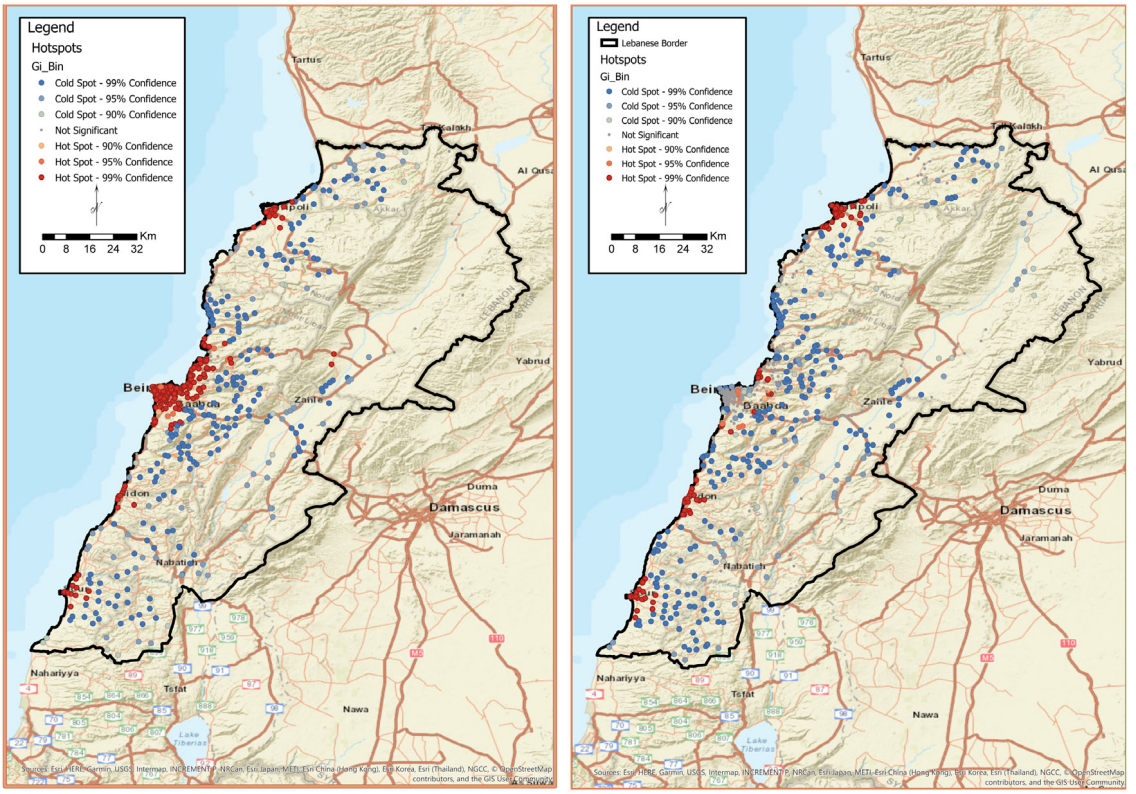
$$V[G] = E[G^2] - E[G]^2 \tag{4}$$

and $E[G^2]$ is defined in Eqs. (8a)–(8k) in Appendix A below.

The n-by-n spatial weight matrix $[w_{ij}]$ is constructed, based on the spatial contiguity matrix. A weight between 1 and 0 is assigned to each feature according to Euclidean distance from its neighbors. To accommodate for the neighbors' influence and avoid sharp boundaries on neighborhood relationships, a Butter-worth filter is applied with a threshold distance as the cutoff for the start of decay. The threshold distance is the Euclidean distance that ensures every feature has at least one neighbor. Features within the threshold distance receive a weight w_{ij} of 1. Beyond the threshold distance, decreasing weight values in accordance with the filter are assigned. Finally, weight values are used to compute the G-score according to Eq. (1).

The z-score and p-value measure the statistical significance to infer whether the null hypothesis should be rejected or not. In our proposed analysis, the null hypothesis states that the accidents' values are randomly distributed, and no spatial clustering is observed. If p-value is statistically small, hence the null hypothesis is rejected and the sign of the z-score becomes relevant. If the z-score value is positive and the observed General G statistic is larger than the expected General G statistic, then this indicates high clustering of events in the study area. If the z-score value is negative and the observed General G statistic is smaller than the expected one, then this indicates that events are dispersed.

Our results reveal highly significant clustering with p value of $0.0005 < 0.001$ and z-score of 20.88. The positive z-score value and the observed General G statistic of 0.166087 exceeding the expected G value, indicate that the spatial distribution of high values in the dataset is more spatially clustered than would be expected if underlying spatial processes were random.



(a) (b)

Fig. 4. (a) and (b) show the spatial seasonal characteristics of accident hot spots in Lebanon [33.8547° N, 35.8623° E] in winter and summer, respectively.

3.2. Hot Spot analysis

The global statistics method of the previous subsection assesses the overall pattern and trend of the data [10]. In this subsection, we are interested to use local statistics tools to assess crashes occurrence frequency feature within the context of neighboring features and compare the local situation to the global situation. Local spatial association have been previously used in various research studies to assess spatial urban modeling [11], to study changes in population and employment [12], and to analysis crime scenes as well [13].

The Getis–Ord G_i^* statistics were used to identify "hot" (high density) and "cold" (low density) areas of car accidents. The Getis–Ord G_i^* local statistic is given as shown in Eq. (5) [9]:

$$G_i^* = \frac{\sum_{j=1}^{n} w_{ij}x_j - \bar{X}\sum_{j=1}^{n} w_{ij}}{S\sqrt{\frac{n\sum_{j=1}^{n} w_{ij}^2 - (\sum_{j=1}^{n} w_{ij})^2}{n-1}}} \tag{5}$$

where x_j is the attribute value for feature j, w_{ij} is the spatial weight between feature i and j, n is equal to the total number of features, $\bar{X}$ and S are the average and standard deviation of all x_j values.

A statistically significant hot (cold) spot has to be characterized by a feature with high (low) value surrounded by other spots with high (low) features as well. The local sum for a feature and its neighbors is compared proportionally to the sum of all features. When the local sum is very different from expected, and when that difference is too large to be the result of random chance, a statistically significant z-score appears [9].

In the proposed hot spot analysis of spatial and temporal patterns of accidents in Lebanon, we focused on the overall yearly crashes occurrence frequency, as well as on seasonal trends of accidents. The winter and summer periods are regarded as temporal variables since these are the two main seasons in Lebanon.

Fig. 3 shows hot spot results for the overall yearly crashes occurrence feature using ArcGIS 10.3. Results reveal clusters of hot spot locations with 99% confidence in Beirut, Tripoli, Sidon, and Tyre cities. One can notice that all identified hot spots are among the Lebanese coastline characterized by high population density and major cities.

Fig. 4a and 4b show the hot spots trends between winter and summer seasons, respectively. Main hot spots in the winter season are located in Beirut. Some other hot spots appear along the coastline in Tripoli and Tyre. Also, some hot spots are

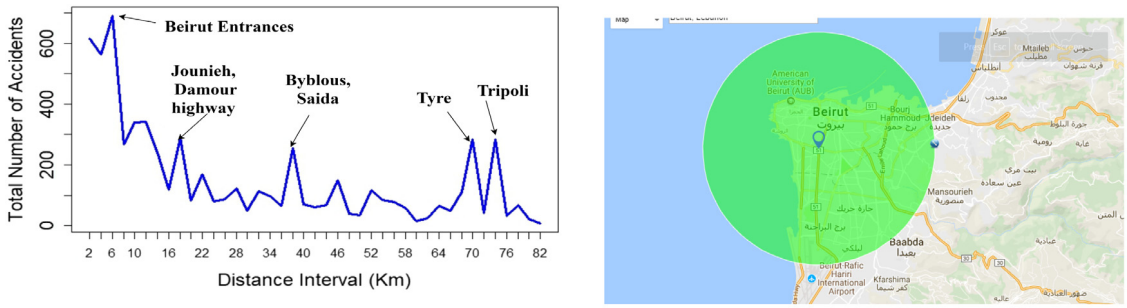


Fig. 5. Figure 5a shows the variation of the total number of accidents with respect to distance and Figure 5b shows 6 km radius from Mousaytbeh.

located in Baalbak area. This can be attributed to the fact that high population density can be found in Beirut during the academic year. Students, professors and staff from the educational sector either live in Beirut or commute everyday to the capital during winter season.

However, in the summer season, hot spots in Beirut are minor as shown in Fig. 4b. This might be caused by the summer break for schools and vacations where most families go to spend their vacation in their hometowns.

In this manner, the presented spatio-temporal patterns empowers the examiner to rapidly and tastefully find measurably acceptable hot spots to ultimately guiding the interested parties to a fruitful administration of movement and decrease of accidents

Finally, after getting the major hot spots from the local G_i^* method, we were interested in observing the variation of the number of accidents as we deviate from the hottest spot, Beirut. The accidents data have been collected over around 160 km radius distance. The center point was taken as the center of the major hot spot area in Beirut which is Msaytbeh with coordinates [33.8836 °S, 35.4955 °N].

As it can be seen in Fig. 5a, the general trend shows that as distance increases from Mousaytbeh, the total number of accidents decrease. However, we observe a major spike at 6 km radius; this observation is best explained by the positions of Beirut entrances (southern, northern, and eastern) within this radius, as shown in Fig. 5b. Beirut entrances inhale massive traffic daily that eventually leads to accidents. The small spikes in the graph are interpreted as accidents in the major cities over the coastline.

4. Hazard vulnerability analysis

Hazard Vulnerability Analysis (HVA) is a very crucial tool for developing national emergency operations plan and disaster management. Given a range of possible alternatives to choose from, HVA allows for a thorough understanding of the risks and the ability to define what could happen in the future. Transportation systems that are disrupted by a hazardous event do represent a critical role in emergency management [14].

There are several ways to measure roads hazardousness, but the most important metric is the number of accidents on that road that reflects how safe is that particular road. As we aim to provide a better driving environment, we propose a novel weight metric to detect the dangerousness of a specific road segment. For future work, we intend to develop this model further to include additional factors such as weather condition, accidents severity index among others.

We defined the road hazard H_i in the following Eq. (6):

$$H_i = \frac{N_i}{L_i * T} \tag{6}$$

where N_i is the number of crashes on road segment i, L_i is the length of road segment i in Km and T is the time span of the whole dataset in days.

For each road segment i with at least one crash record, we calculated road hazard H_i and we normalized all road hazard values to a scale from zero to one according to Eq. (7):

$$Z_i = \frac{H_i - H_{min}}{H_{max} - H_{min}} \tag{7}$$

Fig. 6 shows the distribution of road segment using their corresponding Z_i values as a color weight. Dark red color corresponds to high Z_i values of the most dangerous roads and dark blue color refers to the least dangerous ones. Fig. 6a is depicted for all Lebanon whereas Fig. 6b focuses on Beirut Area. The focus on Beirut Area, with accordance to spatial zooming theory, allows for better hazard analysis to prioritize the high-density zones for intervention efforts [15].

Given the normalized road hazard metric Z_i previously defined, we intended to estimate the hazard metric of other roads segments with missing accidents records. To infer the segments weights based on neighboring segments, we created a matrix

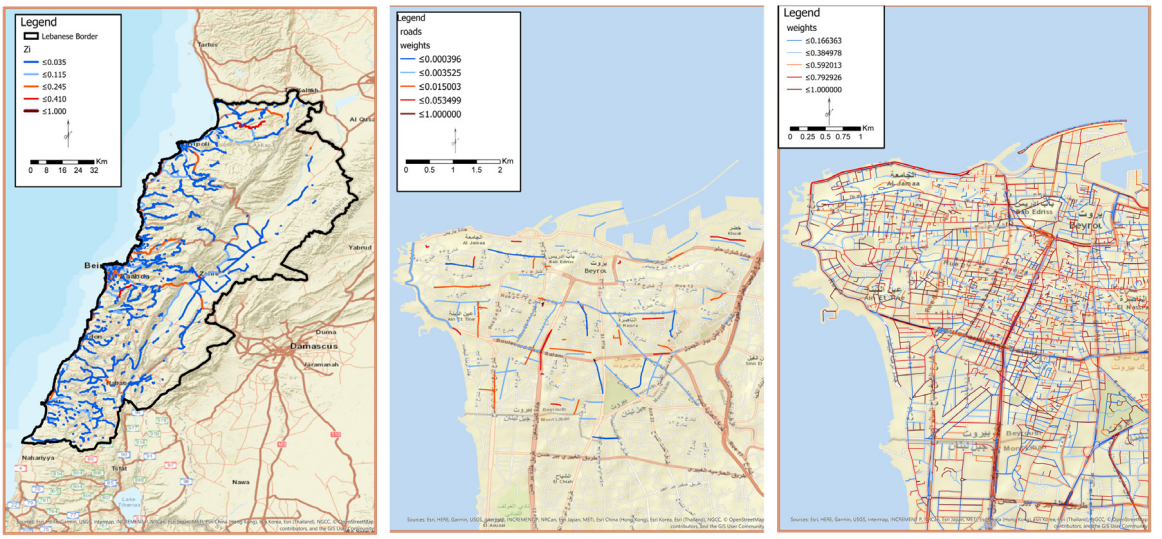


Fig. 6. Figures 6a and 6b show the calculated road hazard map for Lebanon [33.8547° N, 35.8623° E] and Beirut [33.8938° N, 35.5018° E], respectively. Figure 6c shows the estimated road hazard map for Beirut.

stating the neighboring for each road using the queen contiguities. Queen's contiguity considers all cell in contact with the central cell as a neighbor and computes the estimated hazard metric $\tilde{Z}_i$ as the average value of neighboring segments metrics.

Fig. 6c shows the roads network after estimating hazard metric $\tilde{Z}_i$. The estimation is influenced by the number of neighbors having actual data records. Hence, as the percentage of recorded accidents increases, the estimation becomes more accurate and reliable.

5. Conclusion

This paper presents insights about Lebanese road accidents for the period from 2015 until 2018 using crowd-sourced social media data. Several interesting findings were reported. Crash types such as motorcycles accidents were found to be correlated with living conditions. A hot spot analysis was carried out to study the spatio-temporal evolution of car crashes in Lebanon.

Moreover, the first hazard vulnerability map for Lebanese road segments was proposed.

In summary, the contribution of this paper is threefold: (*i*) An analysis of the spatial distribution of car accidents (from 2015 until 2018) based on defined crash types is presented. Correlation conclusions are made between motorcycles accidents and living conditions. (*ii*) The first hot spot analysis map of Lebanese crashes, to the best of our knowledge, is presented where major hot areas were found to be located across the coastline characterized by high population density and major cities. Seasonal trends were clearly observed: during winter season, hot spots are mainly clustered around Beirut Greater Area while in the summer season, hot spots move toward the north and the south of the country. This can be attributed to the summer break where most families leave the capital to spend their vacations. (*iii*) Finally, the most important contribution of this work is to present the first risk assessment and hazard vulnerability map for Lebanese roads. A road hazard index was proposed and used to measure road segments hazardousness. The resultant map can be used as input in various future applications to reduce the high risk of death and injuries induced by vehicle accidents and enable stakeholders to employ more reliable and robust safety measures.

Acknowledgment

This project has been funded with support from National Council for Scientific Research in Lebanon.

Appendix A

$E[G^2]$ mathematical derivation is detailed as shown in Eqs. (8a)–(8k):

$$E[G^2] = \frac{A+B}{C} \tag{8a}$$

$$A = D_0\Big(\sum_{i=1}^{n} x_i^2\Big)^2 + D_1 \sum_{i=1}^{n} x_i^4 + D_2\Big(\sum_{i=1}^{n} x_i\Big)^2 \sum_{i=1}^{n} x_i^2 \tag{8b}$$

$$B = D_3 \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i^3 + D_4\Big(\sum_{i=1}^{n} x_i\Big)^4 \tag{8c}$$

$$C = \Big[\Big(\sum_{i=1}^{n} x_i\Big)^2 - \sum_{i=1}^{n} x_i^2\Big]^2 \times n(n-1)(n-2)(n-3) \tag{8d}$$

$$D_0 = (n^2 - 3n + 3)S1 - nS_2 + 3\Big(\sum_{i=1}^{n}\sum_{j=1, i\neq i}^{n} w_{ij}\Big)^2 \tag{8e}$$

$$D_1 = -\Big[(n^2 - n)S1 - 2nS_2 + 6\Big(\sum_{i=1}^{n}\sum_{j=1, i\neq i}^{n} w_{i,j}\Big)^2\Big] \tag{8f}$$

$$D_2 = -\Big[2nS_1 - (n+3)S_2 + 6\Big(\sum_{i=1}^{n}\sum_{j=1, i\neq i}^{n} w_{i,j}\Big)^2\Big] \tag{8g}$$

$$D_3 = 4(n-1)S_1 - 2(n+1)S_2 + 8\Big(\sum_{i=1}^{n}\sum_{j=1, i\neq i}^{n} w_{i,j}\Big)^2 \tag{8h}$$

$$S_1 = (1/2)\sum_{i=1}^{n}\sum_{j=1, i\neq j}^{n} (w_{i,j} + w_{j,i})^2 \tag{8i}$$

$$S_2 = \sum_{i=1}^{n}\Big(\sum_{j=1, i\neq j}^{n} w_{i,j} + \sum_{j=1}^{n} w_{i,j}\Big)^2 \tag{8j}$$

$$S_2 = \sum_{i=1}^{n}\Big(\sum_{j=1, i\neq j}^{n} w_{i,j} + \sum_{j=1}^{n} w_{i,j}\Big)^2 \tag{8k}$$

References

[1] World Health Organization, in: Global Status Report on Road Safety 2015, Geneva, Switzerland, 2015.
[2] World Health Organization, in: 2nd UN Stakeholders Forum on Road Safety, 2007.
[3] A.J. Ghandour, M. Lovallo, L. Telesca, Time-clustering behavior and cycles in the time dynamics of car accident sequences in Lebanon, Physica A 516 (2019) 178–184.
[4] E. Choueiri, G. Choueiri, B. Choueir, An overview of road safety in Lebanon with particular attention to nonurban roads, in: Advances in Transportation Studies, vol. XI, 2007, Section B.
[5] The Beirut Urban Transport Project, BUTP Preparatory Study, Consultancy Study, World Bank.2003,
[6] D. Dai, E. Taquechel, J. Steward, S. Strasser, The impact of built environment on pedestrian crashes and the identification of crash clusters on an urban university campus, West. J. Emerg. Med. 11 (3) (2010) 294–301.
[7] E. Petridou, M. Moustaki, Human factors in the causation of road traffic crashes, Eur. J. Epidemiol. 16 (9) (2000) 819–826.
[8] Fawaz Mona, Isabelle Peillen, The case of Beirut, Lebanon, in: Understanding Slums: Case Studies for the Global Report on Human Settlements, 2003.
[9] ESRI, ArcGIS Desktop: Release 10.3 Redlands, Environmental Systems Research Institute, CA, 2014.
[10] Arthur Getis, J.K. Ord, The analysis of spatial association by use of distance statistics, Geogr. Anal. 24 (3) (1992).
[11] J. Abed, I. Kaysi, Identifying urban boundaries: Application of remote sensing and geographic information system technologies, Can. J. Civil Eng. 30 (2003) 992–999.
[12] C. Baumont, C. Ertur, J. Le Gallo, Spatial analysis of employment and population density: The case of the agglomeration of Dijon, 1999, Geogr. Anal. 36 (2004) 146–176.
[13] S. Khalid, F. Shoaib, T. Qian, Y. Rui, A.I. Bari, M. Sajjad, J. Wang, Network constrained spatio-temporal hotspot mapping of crimes in Faisalabad, Appl. Spat. Anal. Policy 11 (3) (2017) 599–622.
[14] T.J. Cova, S. Conger, Transportation hazards, in: M. Kutz (Ed.), HandBook of Transportation Engineering, McGraw Hill, New York, 2004, pp. 17.1–17.24.
[15] C. Plug, J. Xia, C. Caulfield, Spatial and temporal visualisation techniques for crash analysis, Accid. Anal. Prev. 43 (6) (2011) 1937–1946.