Physica A 545 (2020) 123534

Contents lists available at ScienceDirect

Physica A

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

Allometric scaling of road accidents using social media crowd-sourced data

Ali J. Ghandour [a,*], Huda Hammoud [b], Mohammad Dimassi [c], Houssam Krayem [d], Jamal Haydar [d], Adam Issa [e]

[a] National Council for Scientific Research, Beirut, Lebanon
[b] Computer and Communication Engineering Department, American University of Beirut, Lebanon
[c] Computer and Communication Engineering Department, Lebanese University, Lebanon
[d] Computer and Communication Engineering Department, Islamic University of Lebanon, Lebanon
[e] Electrical and Computer Engineering Department, University of Toronto, Canada

ARTICLE INFO

Article history:
Received 2 February 2019
Received in revised form 27 May 2019
Available online 22 November 2019

Keywords:
Car accidents
Allometric scaling
Road safety
Crowd-sourcing
Seasonality trends

ABSTRACT

Traffic accidents in Lebanon are constantly harvesting lives, dramatically changing others, and traumatizing those of their beloved ones. Due to the lack of statutory authority in charge of collecting and reporting accident related data, the Lebanese Road Accident Platform (LRAP) is proposed in this work as a real-time online platform to collect crash events from social media. LRAP allows for autonomous data collection, classification and visualization without human intervention, and aims to help the authorities in laying down the appropriate measures for traffic accidents prevention. After being in production for the last four years, the data extracted from LRAP was used to study the allometric scaling of accidents with respect to different parameters such as district area, population size per district and road network length. Such approach offers a new perspective on traffic accidents' scaling and behavior as a living organism as cities grow. A seasonality trend analysis is also provided to analyze temporal clustering patterns in crash occurrence.

1. Introduction

Traffic accidents harvest more than five hundred lives a year in Lebanon, resulting in more than 1.5% cost of the national GDP [1].

Despite the existence of a new Lebanese traffic law that became effective in April 2015, official reports published by authorities still lack advanced spatio-temporal information needed to analyze and assess road conditions, and rather focus on aggregate statistical figures such as number of accidents, number of casualties and fatalities. This is mainly due to the absence of an automated accidents reporting system. Until today, accident reports are still hand written and not properly stored and digitized. In this work, we present the Lebanese Road Accidents Platform (LRAP), a new platform for collecting and analyzing traffic accidents using social media based crowdsourcing. The proposed platform comes as a necessity to collect representative time series of road accidents, with spatial dimension, due to the fact that no statutory authority is constantly collecting and publishing these data.

* Corresponding author.
E-mail addresses: aghandour@cnrs.edu.lb (A.J. Ghandour), hah57@mail.aub.edu (H. Hammoud), md22.dimassi@gmail.com (M. Dimassi), krayem.hossam@gmail.com (H. Krayem), jamal.haydar@iul.edu.lb (J. Haydar), adam.issa@mail.utoronto.ca (A. Issa).

https://doi.org/10.1016/j.physa.2019.123534
0378-4371/

LRAP is a novel platform to mine road accidents related data from various social media sources and make it available in real-time to policy makers, various stakeholders and researchers in the field. Collected data would be useful to localize crashes, find their causes, and more importantly, predict their occurrence given the involved agents and parameters, in addition to various other applications. Our work presents the first real-time interactive and openly available spatio-temporal accidents database of Lebanon. The data collected from this work has proven itself to be credible as it is has been used in multiple research projects so far [2–5]. The proposed methodology is expected to be useful for other developing countries who lack official accidents monitoring and reporting systems.

This manuscript's contribution is three-folds: (*i*) devise a platform to collect real-time data pertaining to traffic accidents from credible social media accounts. Data collection, classification and visualization is done autonomously and in real-time without human intervention. (*ii*) Scaling analysis on collected data with respect to different parameters such as district area, population size and road network length. Results show that accidents tend to follow the same scaling factor of an organism's network making accidents act as living organism in behavior. (*iii*) Temporal analysis reveals seasonal, weekly and daily trends in road crashes clustering.

The rest of this paper is organized as follows: in Section 2, we review existing relevant research in the literature. Section 3 introduces the proposed Lebanese Road Accident Platform (LRAP). Accidents scaling analysis with various parameters is studied in Section 4. Temporal analysis and distribution of road accidents is provided in Section 5. Finally, Section 6 presents the conclusion.

2. Literature review

Crowdsourcing-based analysis for urban development has benefited from the ubiquity of computing in the modern era. Social media, blogs, mobile applications and sensors are typical sources of data used in this kind of research, leading to speed estimations as in [6], virtual crowd simulation [7], proposing solutions to traffic congestions [8], as well as traffic accidents detection [9] and prediction [10].

In particular, mining social media is the cheapest option for crowdsourcing. The work in [9] for instance, collects data from Twitter accounts about traffic accidents in Pittsburgh and Philadelphia Metropolitan Areas, USA.

Authors in [11] propose a deep learning method to detect accident related events from social media data in both Northern Virginia and New York City. The suggested schema relies on geo-tagged tweets, a feature rarely used by Twitter users in developing countries. Also, the work in [8] addresses the congestion problem in some cities in India, and proposes an architecture to solve this issue. The work in [12] measures incidents response by building a text classifier from crowdsourced twitter data. The system was able to find information about incidents that are not available in the official data set such as gas shortage, crowdsourced traffic and closure conditions. The system is built to detect all disaster related incidents and not specifically road accidents.

Other research such as [13] used Twitter feeds along with the "511 incident data feed" to acquire real time accident information such as time and location which is inputed to a virtual sensor API to get travel time at the accident's location. It mainly aims at the representation of travel time variation during and after incidents for traffic management purposes.

Moreover, Twitter data for accident detection was also used in [14]. Originating from the idea that personal tweets can provide new information not provided in the official accidents reporting system, and that a personal tweet is more likely to report an event that has just happened than an organizational account, the authors labeled the tweets as organizational and personal tweets and used different dictionaries for each. The raw data was processed to identify the relevancy and account types of the tweets, the dictionaries were derived, and classifications were performed accordingly. Personal tweets are useful for extraction of new information. However, mining non-trusted social media accounts can sometimes be risky and may yield to inaccurate information and conclusions.

Finally, researchers in [15] used Twitter data to analyze the human influence on road accidents by extracting venue type information from the Twitter check-in data and comparing its spatial distribution to that of the crash data generated by the official traffic recording system.

As opposed to existing research in the literature, our system parses the location from the text and does not rely on geo-tagged tweets, as geo-tagged tweets are not commonly used in Lebanon and other developing countries.

3. Lebanese road accident platform

Lebanese Road Accident Platform (LRAP), a platform to gather crash reports is proposed in this work by employing Natural Language Processing (NLP) techniques to extract accident event features, mainly: (*i*) Accident Time, (*ii*) Location, (*iii*) Injuries Number, (*iv*) Victims Number, (*v*) Person Type (also referred to as "Who Were" by the National Highway Traffic Safety Administration, NHTSA) and (*vi*) Vehicle Type ("That Were").

Dictionaries are built to identify accident related tweets and posts, classify them by types, and retrieve the number of casualties and fatalities in addition to their spatio-temporal localization. Then, traffic accidents, referred to in the following as *Events*, are stored in a geodatabase along with their above mentioned features. This facilitates events visualization and statistical analysis, paving the way for insightful conclusions. Although some features for a specific Event might be missing, this is not a major issue since our interest is to get a time series for the Events along with its spatial dimension and other features. Fine-grained data is not the goal of this work, but rather a sense of the general trends of accidents in Lebanon.

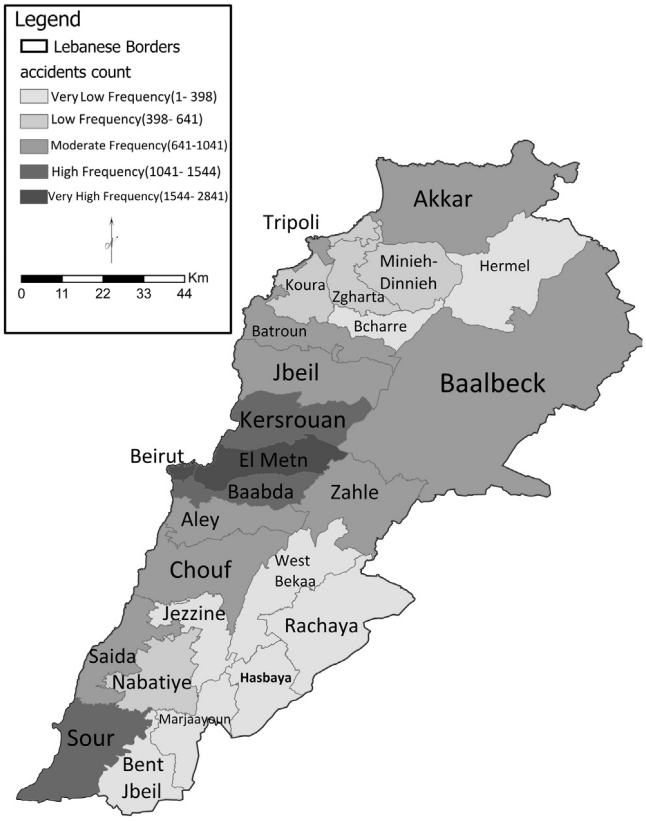


Fig. 1. Distribution of Accident Occurrence Frequency by District [33.8547° N, 35.8623° E].

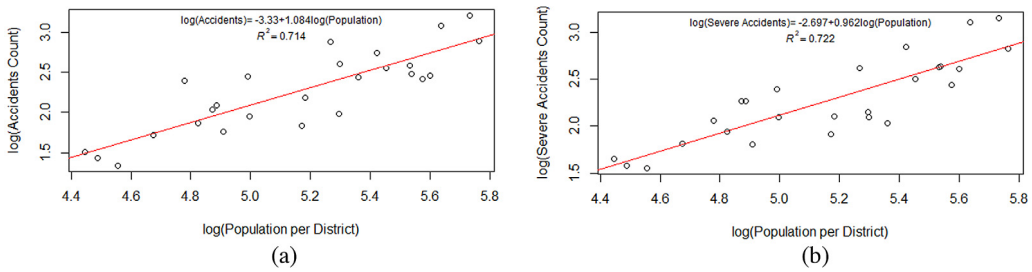


Fig. 2. Fig. 2(a) shows total accidents scaling with population per district in thousands and Fig. 2(b) shows severe accidents scaling with population per district in thousands.

The proposed platform has been in production for four years and reported data spans the period from February 2015 until February 2019.

The main challenges handled towards the full realization of this platform can be summarized as follows: (*i*) no unique sources publish comprehensive data, and thus several social media account were consulted and then merged and fused to avoid multiple counting. (*ii*) Data is published mostly in Arabic, but also in English in some cases which added additional complication to the process, especially that data mining packages that support Arabic language are not fully mature yet. (*iii*) Lack of unified publishing format across monitored social media accounts. In fact, some of these accounts do not follow a standard publishing format at all. (*iv*) Collected tweets and posts are not geotagged. (*v*) Spelling errors made by the publisher.

It is worthy to mention that the proposed platform is of potential use in other developing countries facing the same or similar issues as the ones outlined above. In fact, the lack of geotagging, as well as the absence of a unique accident reporting agency, are typical problems in developing countries. In particular, Arab countries can benefit from LRAP since it is originally designed to parse collected data in Arabic. However, the extension to non-Arabic dictionaries can be easily achieved to make it also useful in countries where non-Arabic dominates interactions on social media.

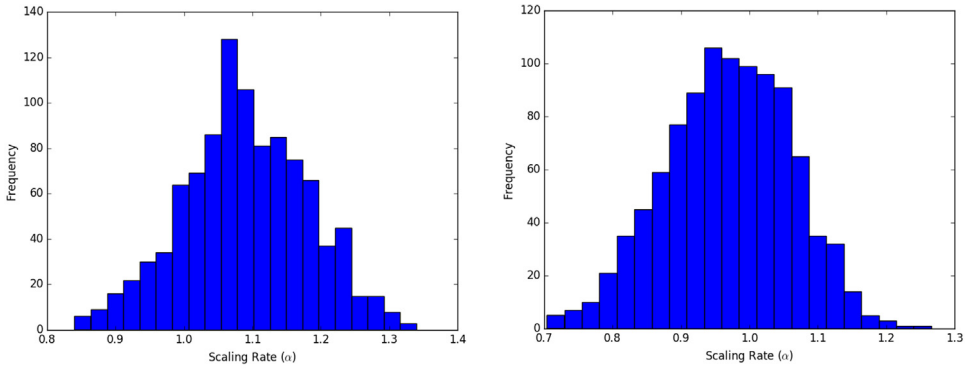


Fig. 3. Figs. 3(a) and 3(b) show the frequency histograms of the scaling rate of total and severe accidents, respectively, with population per district corresponding to 1000 random shuffles.

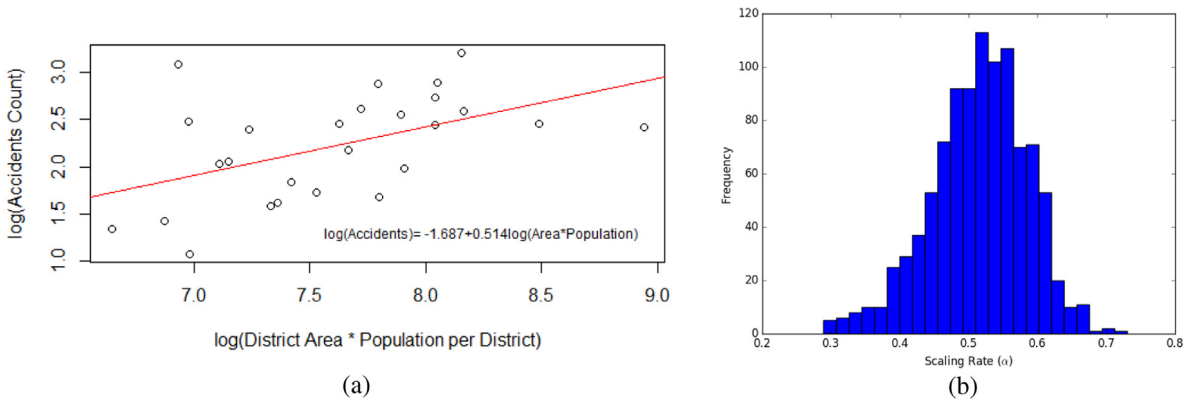


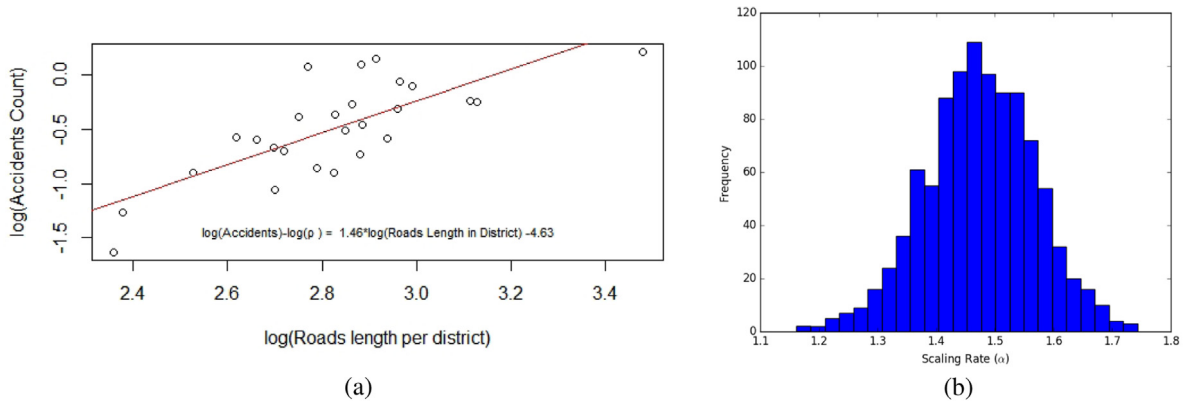
Fig. 4. Fig. 4(a) shows the scaling of accidents rate with district area and population size per district and Fig. 4(b) shows the corresponding scaling rate frequency histogram after 1000 random shuffles.

4. Accidents scaling analysis

Lebanese territory is administratively divided into 26 districts. To account for the zone-level crash risk factors, we first investigated the distribution of crash frequencies over districts to model potential spatial patterns. Fig. 1 shows a representation of crash frequencies associated with different districts in Lebanon. Results reveal a major condensation of accident frequency over Beirut and neighboring districts, also known as Greater Beirut Area. The capital Beirut acts as the center of crash aggregation, and crash frequencies drop the further away from the capital. The high density of crash occurrence in Beirut can be attributed to the fact that it is the centralized capital of the country, with over-population that exceeds 1 million inhabitants, in addition to the increasing number of cars and a quasi-absent public transportation system. Accidents were also more frequent at the coastal line in main arteries where most urban activities take place. Coastal cities such as Tripoli, Sidon, Byblos and Tyre have high traffic volume and are directly connected to the major roads in Lebanon.

To study the scaling effect of car accidents with the population size per district, which is an important factor for strategic planning, we performed linear regression with a log–log transformation on both total accidents count and severe accidents (casualties incurred) count versus population size per district. Figs. 2(a) and 2(b) show that our models are significant with a p-value of $3.47 \times 10^{-8} > 0.05$ and an adjusted R^2 of 0.714 for all accidents count model. As for the severe accidents model, we got a p-value of $2.38 \times 10^{-8} > 0.05$ and an adjusted R^2 of 0.722. Using log–log transformation, we can deduce that for 1% increase in the population there is a 1.08% increase in the number of total accidents and 0.96% increase in the severe accidents. Total accidents rate shows a positive linear increase with the population size while severe accidents rate exhibits a lower increase.

The linear scaling of total accidents count versus population per district ($\alpha = 1.080 \pm 0.136$), shown in Fig. 2(a), is in accordance with results witnessed in USA as reported in [16]. Although different countries have different driving conditions, environmental factors, vehicle conditions, road infrastructure, traffic patterns, and spatial factors, it seems that the scaling of accidents with population follows a universal trend regardless of the geographical location.



(a) (b)

Fig. 5. Fig. 5(a) shows the scaling of accidents rate with roads length and Fig. 5(b) shows the corresponding scaling rate frequency histogram after 1000 random shuffles.

The scaling rate of severe accidents (defined as accidents including casualties) with population per district in Lebanon ($\alpha = 0.960 \pm 0.118$) shown in Fig. 2(b) is found to be exceeding results in the USA. This might be due to various factors including vehicle fleet condition, post-crash response and seat belt use rates among others. The fleets in Lebanon for instance are relatively old with 71% older than 10 years [17]. The old nature of the fleets enhances the severity of accidents since old vehicles lack the advanced safety features that protect drivers.

For better statistical consistency of the results, 1000 random shuffles were applied while calculating the scaling rate (alpha slope) previously reported in Fig. 2 for total accidents with population and severe accidents with population. The corresponding frequency histograms of the results obtained from the shuffles are presented in Figs. 3(a) and 3(b), respectively.

The mean value of the frequency histogram of the scaling rate of total accidents with population per district shown in Fig. 3(a) is 1.089 which is in accordance with the 1.08 scaling rate reported above. Moreover, The mean value of the frequency histogram of the scaling rate of severe accidents with population per district shown in Fig. 3(b) is 0.968 which is in accordance with the 0.96 scaling rate reported above.

To push this analysis further, we studied accidents scaling in relationship to the district area and the population size per district. This is the first time this relation is investigated as far as the authors know. Motivation for this work grew out from [18] which suggests a metabolic scaling theory that compares the road network to the cardiovascular network. Taking into account that advances in biology showed a compelling non-linear relationship between the capacities of networks and sizes of systems in which they are set, researchers in [18] propose a non-linear relation between the road area network, the city area and population size as shown in Eq. (1):

$$A_{roads} \propto (NA_{City})^{1/2} \tag{1}$$

where A_{roads} and A_{city} are the road network and city areas in Km^2, respectively, and N is the population size in thousands.

Fig. 4(a) uses a log–log transformation with linear regression to visualize the relation between accidents count with respect to district area and population size per district. The regression produced a coefficient of ($\alpha = 0.51 \pm 0.18$).

Thus, following the same analogy in [18], we were able to show that the relation between accidents, district area and population size per district follow a similar non-linear relation as described in Eq. (2):

$$Accidents\_Count \propto (NA_{district})^{1/2} \tag{2}$$

where $A_{district}$ is the district area in Km^2 and N is the population size per district in thousands.

Hence, it is found that the relation of accidents count with respect to the district area times population size follows the square root curve;i.e, as we increase the x-dimension, the y-dimension increases with an ever-decreasing rate. The increase rate converges to zero at infinity. Accordingly, the rate of increase in accidents will decrease with the increase of the district area and/or population.

The allometric scaling analysis presented in Eqs. (1) and (2) and Fig. 4(a) has shown that accidents tend to follow the same scaling factor of an organism's network, making accidents a living organism in behavior. This analogy might give us an insight about the expansion, patterns, and future scaling of accidents with city expansion.

Once again, 1000 random shuffles were applied while calculating the scaling rate (alpha slope) for total accidents with district area and population. The mean value of the corresponding frequency histogram shown in Fig. 4(b) is 0.514 which is aligned with the 0.51 scaling rate reported above.

Finally, along the lines of the analogy given above, we study accidents in relation to road networks and district density. In this case, we draw the analogy from the relation of the vessels delivery network to the organism defined in [18] as follows:

$$V_{net} \propto \rho V_{org}^{4/3} \tag{3}$$

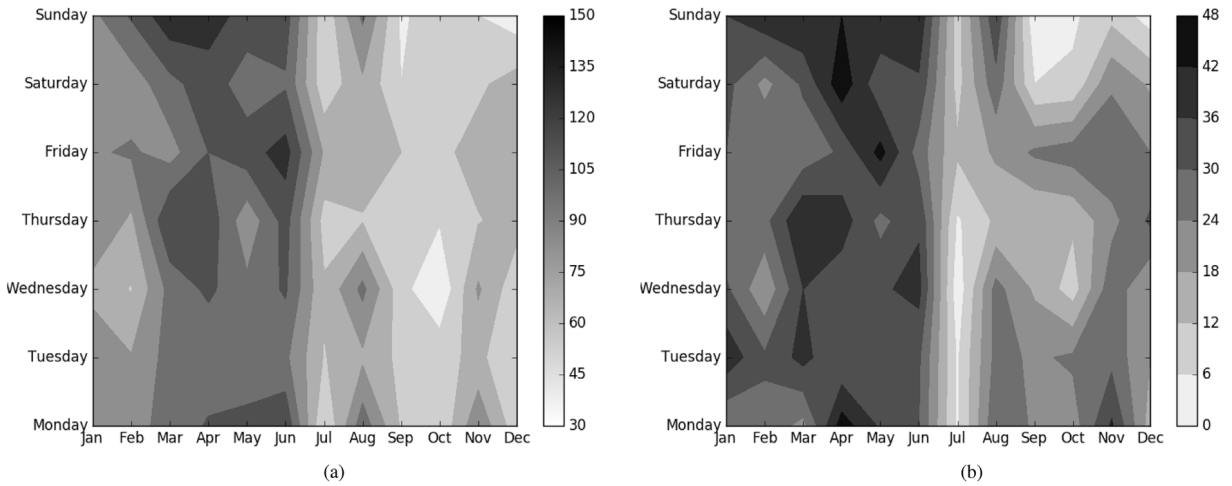


Fig. 6. Fig. 6(a) shows the density plot of the frequency of accidents by month and day of the week and Fig. 6(b) shows the density plot of the mean of accidents by month and day of the week corresponding to 1000 random shuffles.

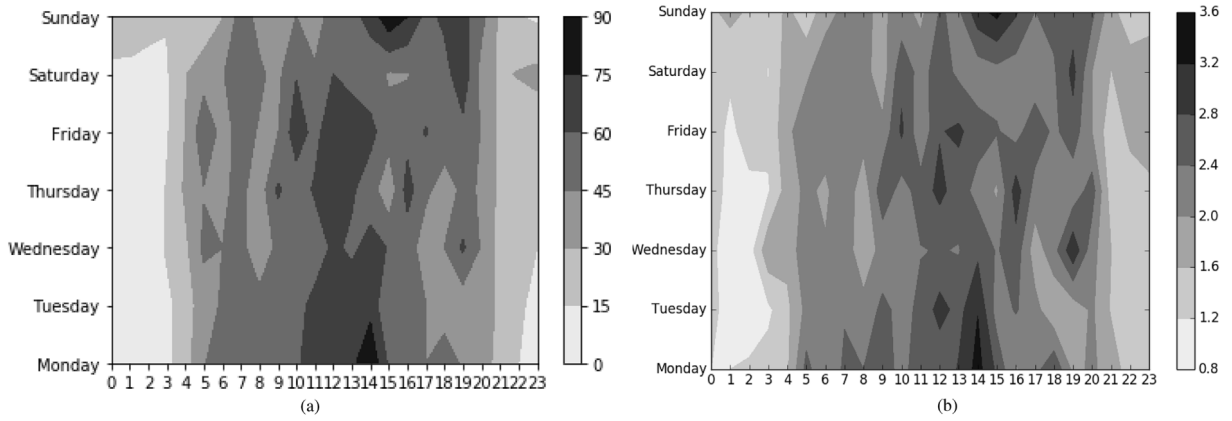


Fig. 7. Fig. 7(a) shows the density plot of the frequency of accidents by hour and day of the week and Fig. 7(b) shows the density plot of the mean of accidents by hour and day of the week corresponding to 1000 random shuffles.

where V_{net} and V_{org} are the volumes of the blood vessels and the organism respectively, and ρ is the cell density.

The exponent in Eq. (3) reflects the 3-dimensional structure which can be generalized to a system in any dimension D to be $(D+1)/D$ [18]. Hence, in a 2-dimensional district's road environment, the exponent would be 3/2. Thus, the relation between accidents count, road network length and district density should be as follows:

$$Accidents\_Count \propto \rho Roads\_Length^{3/2} \tag{4}$$

where ρ is the district density defined as the district population in thousands divided by the district area in Km^2, and *Roads_Length* is the total roads length in a district in *Km*.

Using regression to study the non-linearity nature of the above relation in Eq. (3), we were able to produce a factor of 1.46 as shown in Fig. 5(a) ($\alpha = 1.46 \pm 0.25$) proving the scaling analogy between the accidents in roads and the blood network in a living organism.

The mean value of the frequency histogram of the scaling rate of accidents with roads length corresponding to 1000 random shuffles shown in Fig. 5(b) is 1.473 which is aligned with the reported scaling rate value of 1.46.

By understanding the properties of the distribution network within an organism, it is possible to recognize an important set of primary constraints on how this organism functions. Thus, the analysis proposed here allows decision making and researchers to focus and cope with the scaling rate at which accidents occur within a certain region of interest.

Table 1
Area, population and roads length per district.

District	Area (in km^2)	Population	Roads Length (in km)
Beirut	19.6	432,645	238.5949885
Tripoli	27.3	345,343	227.8129272
Baabda	194	579,382	865.8641449
El-Meten	263	541,316	981.4861493
Bint-Jbeil	264	99,224	757.8195952
Sidon	274	283,563	769.5978902
Aley	264	199,324	729.8063265
Chouf	481	228,722	1304.418407
Koura	173	74,374	458.2282262
Marjeyoun	265	80,799	667.9888128
Keserwan	336	184,748	820.8779486
Nabatieh	304	152,042	707.6295365
Zgharta	182	76,964	415.8223551
Tyre	414	263,861	920.6335163
Zahle	425	341,552	914.0275986
Bcharre	158	27,881	336.9522659
Jezzine	242	30,771	497.766621
Batroun	287	59,977	588.3908669
Hasbaya	265	36,012	500.1981005
Byblos	430	98,041	762.9430414
Akkar	778	397,823	1348.164088
Western Beqaa	425	147,995	615.4664055
Rachaiya	485	47,221	673.8850441
Baalbek	2319	375,697	3015.074625
El Miniteh-Danniyeh	409	197,897	523.1986424
Hermel	506	66,698	562.3924548

5. Seasonality trends

Traffic density patterns on hourly, weekly, and monthly scales can be obtained from density plots. They are used to analyze accident peaks in terms of the total number of accidents in certain time periods.

Fig. 6(a) shows a seasonal trend in the number of car accidents. Namely, the winter season has the highest concentration. This is due to the relatively worse road conditions during winter compared to other seasons. In fact, the Sunday peak in March and April is correlated with the heavy rain and bad weather the country witnesses during that month. Future studies will focus on the correlation between weather conditions and road accidents. The summer Friday peak on June is mainly related to summer vacation. Following the stressful period of exams, people tend to go out more, especially in a good weather, leading to more traffic flow in the roads causing more traffic accidents. Fig. 6(b), which shows the density plot of the mean of accidents count by month and day of the week corresponding to 1000 random shuffles, confirms the seasonal trend and the peak observed during winter season.

Fig. 7(a) shows that a large percentage of the accidents is clustered in the time interval between 10am and 5pm, the usual work time. A peak can be observed on Sunday afternoon, which can be interpreted by the weekly migration back to Beirut. After spending the weekend in the mountains, huge traffic is witnessed on Sunday afternoon where citizens return back to the capital Beirut.

During week days, denser frequency of accidents is mainly witnessed between noon and 2pm when people start leaving work and students leave schools and universities. The peak can be observed at 2pm on Monday because it is the start of the work week, where lots of tasks pile up. Fig. 7(b), which shows the density plot of the mean of accidents count by hour and day of the week corresponding to 1000 random shuffles, confirms previous peak observed on Sunday afternoon and Monday.

Results reported in this section and the previous one constitute a new case study for Lebanon showing that road accidents tend to follow some universal trends. The proposed LRAP platform does not only help the authorities tackle the most pressing aspects of the problem, but also provide a schema for prediction and future planning. Scaling rates and spatio-temporal distribution of road accidents are crucial inputs for decision makers. Emergency or first responder for instance would benefit from the spatial localization of accidents density to deploy centers in such a way that maximal efficiency is reached with minimal commute times.

The proposed work helps in characterizing the dynamics and the allometric scaling of the crash occurrence problem where accidents tend to behave as living organisms. Findings can be easily extended to the Arab region and other developing countries that share similar ecosystem.

6. Conclusion

Lebanese Road Accident Platform (LRAP), a new platform to gather crash reports using crowd-sourcing from social media is proposed in this work to extract accident event features. LRAP has been in production for four years. Reported

results provide insight on the distribution of accidents across Lebanese districts and their allometric scaling with the population size per district, revealing some universal trend. Findings reveal the non-linearity in the scaling behavior of accidents count with district area, population per district and road network length. Accidents are interpreted as a living organism due to their similar behavior in reacting to system capacity and size changes. In addition, seasonality analysis showed the trends exhibited in the temporal distribution and clustering of road car crashes in Lebanon. Future work will focus on using neural network and machine learning techniques to further investigate collected data.

Acknowledgment

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

Appendix

See Table 1.

References

[1] Elias Choueiri, Georges Choueiri, Bernard Choueiri, An overview of road safety in Lebanon with particular attention to non urban roads, in: Advances in Transportation Studies, Vol. XI, Section B, 2007.
[2] N. Mouawad, Speed Traps Spatio-Temporal Optimal Allocation on Lebanese Highways (Master thesis), Lebanese University, 2018, Sept. 2016.
[3] R. Naja, N. Mouawad, A.J. Ghandour, K. Fawaz, Speed trap optimal patrolling: STOP playing stackelberg security games, Wirel. Pers. Commun. 98 (4) (2018) 3563–3582.
[4] Ali J. Ghandour, Michele Lovallo, Luciano Telesca, Time-clustering behavior and cycles in the time dynamics of car accident sequences in Lebanon, Physica A 516 (2019) 178–184.
[5] Ali J. Ghandour, Huda Hammoud, Luciano Telesca, Transportation hazard spatial analysis using crowd-sourced social network data, Physica A 520 (2019) 309–316.
[6] H. Hu, G. Li, Z. Bao, Y. Cui, J. Feng, Crowdsourcing-based real-time urban traffic speed estimation: From trends to speeds, in: Data Engineering, ICDE, 2016 IEEE 32nd International Conference on, 2016, pp. 883–894.
[7] M. Pouke, J. Goncalves, D. Ferreira, V. Kostakos, Practical simulation of virtual crowds using points of interest, Comput. Environ. Urban Syst. 57 (2016) 118–129.
[8] T. Roopa, A.N. Iyer, S. Rangaswamy, Crotis-crowdsourcing based traffic information system, in: Big Data, BigData Congress, 2013 IEEE International Congress on, 2013, pp. 271–277.
[9] Y. Gu, Z.S. Qian, F. Chen, From Twitter to detector: Real-time traffic incident detection using social media data, Transp. Res. C 67 (2016) 321–342.
[10] H. Park, A. Haghani, Real-time prediction of secondary incident occurrences using vehicle probe data, Transp. Res. C 70 (2016) 69–85.
[11] Z. Zhang, Q. He, J. Gao, M. Ni, A deep learning approach for detecting traffic accidents from social media data, Transp. Res. C 86 (2018) 580–596.
[12] A. Kurkcu, F. Zuo, J. Gao, E. Morgul, K. Ozbay, Crowdsourcing incident information for disaster response using Twitter, in: Transportation Research Board 96th Annual Meeting, Washington DC, United States, 2017.
[13] A. Kurkcu, E.F. Morgul, K. Ozbay, Extended implementation method for virtual sensors, Transp. Res. Rec. J. Transp. Res. Board 2528 (2015) 27–37.
[14] M.A. Yazici, S. Mudigonda, C. Kamga, Incident detection through Twitter, Transp. Res. Rec. J. Transp. Res. Board 2643 (2017) 121–128.
[15] J. Bao, P. Liu, H. Yu, C. Xu, Incorporating twitter-based human activity information in spatial analysis of crashes in urban areas, Accid. Anal. Prev. 106 (2017) 358–369.
[16] Abdel-Aty Huang, Darwich, County-level crash risk analysis in florida Bayesian spatial modeling, Transp. Res. Rec. J. Transp. Res. Board (2010).
[17] MoE/UNDP/GEF, National Greenhouse Gas Inventory Report and Mitigation Analysis for the Transport Sector in Lebanon. Beirut, Lebanon, 2015.
[18] H. Samaniego, M. Moses, Cities as organisms: Allometric scaling of urban road networks, J. Transp. Land Use (2008).