sustainability

MDPI

Article

Analyzing the Relationship Between User Feedback and Traffic Accidents Through Crowdsourced Data

Jinguk Kim [1], Woohoon Jeon [1] and Seoungbum Kim [2,*]

[1] Department of Highway and Transportation Research, Korea Institute of Civil Engineering and Building Technology, 283, Goyang-daero, Ilsanseo-gu, Goyang-si 10223, Republic of Korea; jingukkim@kict.re.kr (J.K.); cwhoon@kict.re.kr (W.J.)

[2] Department of Urban Engineering, Engineering Research Institute, Gyeongsang National University, 501, Jinju-daero, Jinju-si 52828, Republic of Korea

* Correspondence: kimsb@gnu.ac.kr; Tel.: +82-55-772-1778

Abstract: Identifying road segments with a high crash incidence is essential for improving road safety. Conventional methods for detecting these segments rely on historical data from various sensors, which may inadequately capture rapidly changing road conditions and emerging hazards. To address these limitations, this study proposes leveraging crowdsourced data alongside historical traffic accident records to identify areas prone to crashes. By integrating real-time public observations and user feedback, the research hypothesizes that traffic accidents are more likely to occur in areas with frequent user-reported feedback. To evaluate this hypothesis, spatial autocorrelation and clustering analyses are conducted on both crowdsourced data and accident records. After defining hotspot areas based on user feedback and fatal accident records, a density analysis is performed on such hotspots. The results indicate that integrating crowdsourced data can complement traditional methods, providing a more dynamic and adaptive framework for identifying and mitigating road-related risks. Furthermore, this study demonstrates that crowdsourced data can serve as a strategic and sustainable resource for enhancing road safety and informing more effective road management practices.

Keywords: driving safety; crowdsourced data; road safety management; environmental factors; hotspot analysis

Citation: Kim, J.; Jeon, W.; Kim, S. Analyzing the Relationship Between User Feedback and Traffic Accidents Through Crowdsourced Data. *Sustainability* **2024**, *16*, 9867. https://doi.org/10.3390/su16229867

Academic Editors: Keshuang Tang, Ashish Bhaskar and Hong Zhu

Received: 6 October 2024
Revised: 9 November 2024
Accepted: 11 November 2024
Published: 12 November 2024

1. Introduction

Over the past decade, traffic fatalities have exhibited a consistently upward trend, correlating with the expansion of road mileage. According to statistics from the Federal Highway Administration (FHWA) [1], road mileage increased by approximately 2.5% between 2013 and 2022, an average annual growth rate of 0.3%. Similarly, the number of registered vehicles has risen at an average rate of 1.5% per year. This expansion of road infrastructure, coupled with increased vehicle usage, has contributed to a rise in traffic fatalities. Notably, the fatality rate per 100 million Vehicle Miles Traveled (VMT) has increased by 23%, rising from 1.10 to 1.35 over the past decade. These trends underscore the growing safety concerns associated with the expansion and utilization of transportation systems. Moreover, similar patterns in road mileage, VMT, and traffic fatalities are observed not only within the European Union but also in developing nations, such as South Korea.

Traffic accidents can be attributed to various human factors, including distracted driving, speeding, and drowsy driving, as well as environmental factors, road conditions, technical issues, weather, road infrastructure, and traffic system defects [2]. In terms of human factors, emerging technologies, particularly autonomous vehicles, are anticipated to play a crucial role in accident prevention. While road management agencies conduct routine inspections to mitigate accidents resulting from hazardous road conditions, inadequately maintained roads continue to incur significant social costs due to the ongoing expansion of road networks, increasing road deterioration, and a shortage of personnel dedicated to road

management. For instance, potholes can cause drivers to lose control, potentially leading to accidents. In 2021, potholes were responsible for 0.8% of road accidents, resulting in 1.4% of fatalities and 0.6% of injuries [3]. Furthermore, road irregularities reduce vehicle speeds by 55% and increase emissions by 2.49% [4].

Although human factors account for the majority of traffic accidents, the contribution of environmental factors is significant; for instance, Ref. [5] highlights that environmental factors are responsible for 34% of all accidents. To address accidents related to environmental conditions, ongoing research is exploring the use of various sensing technologies, drones, and artificial intelligence for effective road and facility management. However, these methods have not yet been widely implemented on actual roads. As mobile phone usage proliferates, crowdsourced data have emerged as a promising and sustainable alternative for addressing traffic accident issues by leveraging public observations and insights. This approach facilitates the real-time collection of detailed, localized information on road conditions, hazards, and traffic patterns that may not be captured by traditional data sources. By involving citizens in reporting and analyzing traffic-related incidents, traffic agencies can gain a more comprehensive understanding of high-risk road segments and their underlying factors. Moreover, engaging the community enhances public awareness and participation in road safety initiatives, thereby contributing to the reduction in traffic accidents and the enhancement of overall road safety. Above all, the sustainability of crowdsourced data in traffic safety lies in its ability to provide ongoing, scalable, and cost-effective insights for improving transportation systems.

Traditional research on identifying high-risk road segments typically employs inductive reasoning based on historical data collected from road sensors. However, the variability in conclusions as new information becomes available represents a drawback in terms of flexibility. This study aims to utilize crowdsourced data as an alternative means of pinpointing areas where accidents are expected to occur due to environmental factors. Rather than relying solely on inductive reasoning, this study introduces a deductive approach that includes the analysis of crowdsourced data as a potential direct factor in accidents. The hypothesis of this study is that "traffic accidents occur in areas where user feedback is frequent", and this hypothesis is tested using crowdsourced data and fatal accident records. Instead of focusing exclusively on a quantitative analysis of the correlation between user feedback and traffic accidents, this study's significance lies in its empirical investigation of how crowdsourced data based on user feedback can be utilized in relation to traffic accidents.

2. Literature Review

Crowdsourcing data is emerging as an extensive yet cost-effective method for gathering traffic-related information [6–8], as it "makes every user an instant social sensor" [9]. This research investigates the correlation between traffic accidents and user feedback, proposing novel applications based on crowdsourced data. To this end, a comprehensive review of existing literature on crowdsourcing in transportation has been conducted. The findings categorize the literature into three distinct groups based on their specific applications in transportation.

The first group of studies utilizes crowdsourced data for incident detection. Salas et al. [10] examined the feasibility of using Twitter for real-time incident detection in the United Kingdom (UK). They proposed a comprehensive methodology that integrates Natural Language Processing (NLP) techniques with a Support Vector Machine (SVM) algorithm to classify public tweets, demonstrating its applicability in identifying traffic-related tweets. Pandhare et al. [11] leveraged tweets related to traffic and accidents to detect road events. In this study, logistic regression and SVM classifiers are applied in conjunction with text-mining techniques to assign appropriate class labels to road events.

Similarly, Zhang et al. [8] employed deep belief network (DBN) and long short-term memory (LSTM) techniques to detect traffic accidents using Twitter-based data. Lu et al. [12] proposed a crowdsourced approach to forecast city-level traffic incidents by collecting social media data on adverse weather and traffic reports. They combined adverse weather reports

and weather-related data from Weibo and tweets to develop a regression model, which demonstrated superior predictive performance in forecasting city-level traffic incidents compared to traditional approaches. Dabiri and Heaslip [13] employed a method known as "bag-of-words" for incident detection, transforming tweets into numerical feature vectors that can be processed by computers. They utilized unsupervised deep learning algorithms for modeling tweets and implemented supervised deep learning architectures. Rettore et al. [14] introduced the Road Data Enrichment (RoDE) framework, which leverages Twitter data to enhance Intelligent Transportation System (ITS) services through Twitter MAPS (T-MAPS) for route planning and Twitter Incident (T-Incident) for event detection. While T-MAPS achieves up to 62% similarity with Google Maps' routes, T-Incident demonstrates over 90% accuracy in identifying traffic events, showcasing its superior performance in incident detection. However, this study also revealed the limitations of crowdsourced data in planning applications. Alkouz et al. [15] presented SNSJam, a system that employs cross-lingual data (English and Arabic) from Twitter and Instagram to detect and predict traffic jams. Experimental findings demonstrated that integrating data streams from multiple languages and platforms significantly improves the accuracy of traffic event detection. Waze data are another significant source of crowdsourced information for incident detection. Amin-Naseri et al. [16] investigated Waze data to identify the characteristics of this social sensor and to provide a comparison with common data sources in traffic management. They empirically demonstrated that crowdsourced data could offer extensive coverage, providing timely reporting while maintaining reasonable geographic accuracy. Recently, several studies [17–19] have developed methodologies to detect road incidents by processing Twitter or Waze data.

The second group of studies utilizes crowdsourcing data to detect or monitor pavement conditions. Monitoring pavement conditions is essential for effective pavement management and maintenance. Traditional methods, such as accelerometers, videos, and laser scanning, are constrained by equipment and labor limitations, which can delay maintenance actions. Recent studies have focused on utilizing Waze data to monitor pavement conditions. Gu et al. [20] employed crowdsourced Waze data to evaluate pavement conditions, proposing Pothole Report Density (PRD) and Weather Report Density (WRD) as surrogate measures. They utilized a geographically weighted random forest (GWRF) model to analyze the relationship between crowdsourced data and the official Pavement Quality Index (PQI), finding that PRD exhibits a high correlation with the PQI. Similarly, Liu et al. [21] demonstrated significant benefits of crowdsourced data in pothole detection. Gu et al. [22] developed a framework for pothole detection and evaluation using reports from the Waze app, employing two spatiotemporal density models: STKDE and ST-DBSCAN. This framework was validated against official pavement maintenance records in Nashville, Tennessee. The study found that crowdsourced reports are capable of accurately identifying existing potholes while also revealing additional potholes that regular patrols may overlook.

The third group of studies employs crowdsourcing data for planning and modeling in the field of transportation. Lin et al. [23] introduced the Topic-Enhanced Gaussian Process Aggregation Model (TEGPAM) to predict road speed using multi-source data, including INRIX data and tweets. They addressed challenges such as location uncertainty, language ambiguity, and data heterogeneity. Liao et al. [24] developed a data fusion framework to compare travel times by car and public transit. They combined multiple data sources to estimate travel times for both modes, including traffic data, transit data, and travel demand estimated from Twitter data. Lin and Li [25] created a traffic accident impact model using crowdsourcing data, categorizing accident-induced congestion into four levels and extracting spatiotemporal features, weather information, and accident details from the crowdsourced data. They trained three classification models and tested them to predict congestion levels and durations. Essien et al. [26] introduced a deep learning model that integrates tweet data with traffic and weather information. Their model demonstrated improved accuracy in traffic flow predictions when tested in Greater Manchester compared

to classical and machine learning models. Janež et al. [27] investigated the potential of crowdsourcing data to supplement or replace conventional vehicle counters, such as inductive loop counters (ILC). In this study, crowdsourced data were collected from Telraam counters, which are low-cost cameras operated by citizens. They applied regression models to compare ILC and Telraam counters across four segments. Another study employed user-generated feedback for real-time decision making in traffic management and planning. Dienstl and Scholz [28] focused on utilizing user-generated feedback, specifically through Volunteered Geographic Information (VGI), in demand-responsive transport systems. By integrating user feedback directly into transport management, the study demonstrated how this approach can provide immediate, actionable data for service improvements. A pilot project in Austria confirmed that citizen-sourced data are an effective tool for real-time decision making in transport systems. Liu and Feng [29] developed a deep learning model to predict speed using crowdsourced police enforcement data, arguing that various underlying factors, such as police enforcement, can influence driving behaviors and should be considered in the development of speed prediction models.

This study distinguishes itself from previous studies in several key aspects. The main difference from existing studies specifically aims to leverage crowdsourced data for safety management instead of traffic management. Existing studies indicate that user feedback in crowdsourced data is employed in various applications, such as incident detection, traffic condition prediction, route planning, and the identification of road conditions such as potholes. The majority of studies utilize crowdsourced data predominantly for traffic management and operations, with relatively few addressing safety management. Although some studies [8,10–19] leverage crowdsourced data to detect accidents or predict traffic conditions resulting from incidents, they typically consider accidents as one of several traffic incidents, rather than focusing on accident detection as the primary objective. Specifically, we discuss the potential applications of crowdsourced data in identifying high-crash locations within the context of accident analysis. Traditionally, traffic conditions (e.g., traffic flow and speed), road characteristics (e.g., road type and number of lanes), and environmental data (e.g., weather conditions) have been employed to identify high-crash locations. In contrast, this study explores whether crowdsourced data can be utilized to classify high-crash locations with greater precision.

Additionally, this research employs a nationally developed application to gather feedback from road users. Previous studies have predominantly relied on commercial applications such as Twitter or Waze to collect significant user feedback for their research. However, as highlighted by prior studies [13,22,30,31], the reliability of such data is often questionable, necessitating further refinement, which incurs significant time and cost. Conversely, this study has developed and utilized a specialized application designed to collect road user feedback specifically for road management. This method is anticipated to offer a comparatively higher level of data reliability while requiring less cost and effort for data refinement.

Finally, during the literature review process, we identified a similar study that compares police crash reports with Waze incident reports [32]. Unlike that study, this research focuses on identifying high-crash areas within a national road network, distinguishing these areas more precisely by utilizing road complaint reports.

3. Methodology

This study aims to conduct a spatial analysis of crowdsourced data alongside fatal accident data to assess the hypothesis that traffic accidents occur in areas with frequent user complaints. To verify this hypothesis, it was essential to determine whether each dataset is randomly distributed or follows a distinct spatial pattern. Thus, the first step of the methodology was to perform a spatial autocorrelation analysis based on the location information of each dataset.

Spatial autocorrelation refers to the degree to which a set of spatial data points is correlated with itself based on their location in space. It measures whether objects or

events that are close to each other in a geographical area are more similar (positive spatial autocorrelation) or dissimilar (negative spatial autocorrelation) than those that are further apart. Figure 1 illustrates hypothetical examples of geographical spaces exhibiting positive and negative autocorrelation.

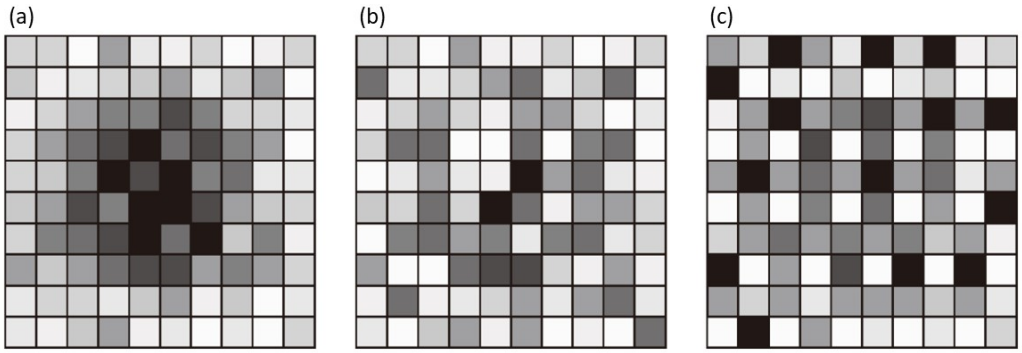

Figure 1. Examples of spatial autocorrelation [33]. (**a**) Positive spatial autocorrelation. (**b**) Spatial randomness. (**c**) Negative spatial autocorrelation.

Spatial autocorrelation is typically measured using statistical indicators such as Moran's I and Geary's C. This study employs local Moran's I, defined by Equations (1) and (2), as it is well-suited for analyzing spatially dispersed events, such as traffic accidents:

$$I_i = \frac{x_i - \overline{x}}{m} \sum_{j=1}^{N} w_{ij}\left(x_j - \overline{x}\right), \tag{1}$$

$$m = \frac{\sum_{i=1}^{N}(x_i - \overline{x})^2}{N}, \tag{2}$$

where

N: number of analysis spaces;
i,j: analysis space (1, 2, 3, . . . *N*, $i \neq j$);
w_{ij}*:* weight between spaces *i* and *j*.

If the distribution of an event exhibits positive spatial autocorrelation, it can be interpreted that data with similar characteristics are geographically clustered together. For instance, if areas with a high incidence of traffic accidents are located close to one another, these regions can be described as exhibiting positive spatial autocorrelation.

After evaluating Local Moran's I to determine spatial autocorrelation for each dataset, hotspot analysis was conducted exclusively on those datasets that exhibit positive spatial autocorrelation. This study employs a density-based clustering technique known as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) for the hotspot analysis. DBSCAN forms clusters based on data density, creating clusters only in areas with a higher concentration of data while treating data in low-density areas as noise and excluding them. A key advantage of this density-based clustering technique is that it does not require the user to pre-specify the number of clusters, making it effective even when cluster densities vary.

Hotspot analysis facilitates the visual exploration of the spatial distribution of multiple events. To evaluate whether the occurrence of one event influences another, this study leverages the results of the hotspot analysis to classify areas where both events occur simultaneously (i.e., where two different hotspots overlap). By calculating the event occurrence density within the identified hotspots for each dataset and comparing the densities between the two classified groups, this study seeks to clarify the relationship between the occurrences of the two events. Figure 2 illustrates the methodology employed in this study.

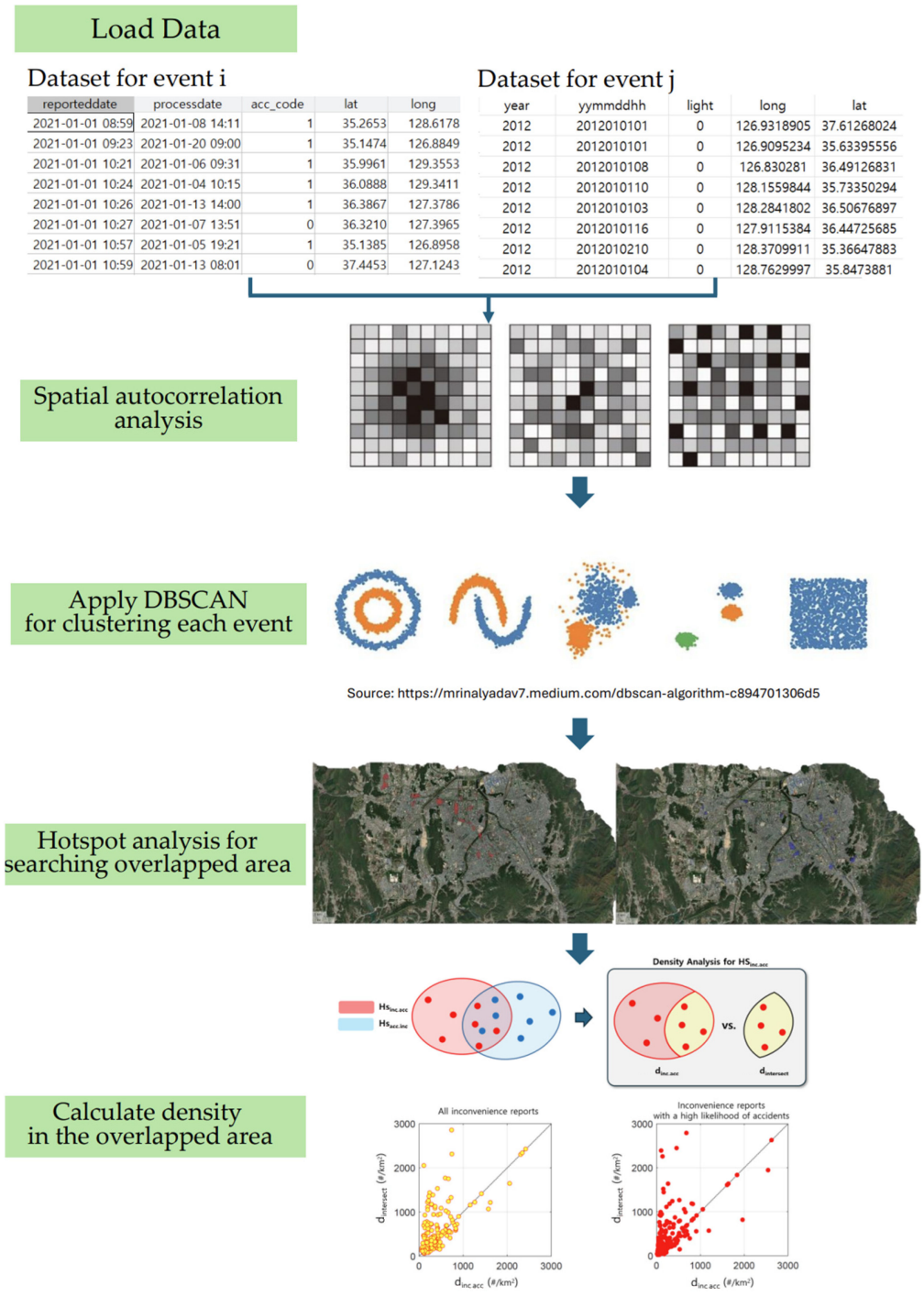

Figure 2. Methodology.

4. Analysis Data and Research Area

To validate the research hypothesis, this study conducts a spatial analysis using two sets of data: crowdsourced data and fatal accident occurrence data. The Ministry of Land, Infrastructure, and Transport in South Korea developed a mobile-based crowdsourcing data collection application, which was launched in 2013. This dataset includes records of inconveniences experienced by road users, as well as follow-up reports on the actions taken to address these issues. The data were collected from 2014 to 2022. Table 1 presents typical examples of each attribute included in the crowdsourced data.

Table 1. Examples of citizen science report.

ID	Date	Inconvenience	Region	Processing Time (h)	Long	Lat
8574	1 April 2014 08:00	stop sign replacement	A	1	126.9	37.6
8584	1 April 2014 09:48	rock on the road	B	6	126.9	35.6
8623	2 April 2014 09:34	roadkill	C	55	126.8	36.5
8639	2 April 2014 14:27	uneven pavement	E	166	128.2	35.7

The inconvenience reports are classified into six categories: poor road surface condition, potholes, roadkill and falling rocks, poor drainage, defective road facilities, and others. The 'others' category encompasses cases where reports are either unspecified or cannot be classified under the five primary categories. The records are collected and classified into 17 regions based on location data. The dataset used in this study contains 65,680 inconvenience reports across these 17 regions for the period from 2014 to 2022.

Fatal accident data were obtained through an open API provided by the Traffic Accident Analysis System (TAAS), operated by the Korea Road Traffic Authority. This dataset was also collected from 2014 to 2022 and includes information on the number of fatalities, injuries, types of accidents, and locations where the accidents occurred. This study utilizes the fatal accident dataset associated with the crowdsourced data, which includes 42,602 fatal accident records. Table 2 presents a sample of the fatal accident data.

Table 2. Examples of fatal accident data.

Date	Fatality	Casualty	Category	Long	Lat
1 January 2014	1	1	While on roadway	126.9	37.6
1 January 2014	1	6	Head-on collision	126.9	35.6
1 January 2014	1	1	Collision with structure	126.8	36.5
1 January 2014	2	2	Side collision	128.2	35.7
1 January 2014	1	1	Road departure and fall	128.3	36.5
1 January 2014	1	1	Collision with structure	127.9	36.4

The analysis area encompasses all regions of South Korea where both crowdsourced data and fatal accident data are available. Figure 3 illustrates the visualization of the extracted location information from each dataset, applying the Lambert conformal conic projection to convert it into a Cartesian coordinate system. The rectangular area in the Figure 3 represents the area of the Seoul metropolitan area.

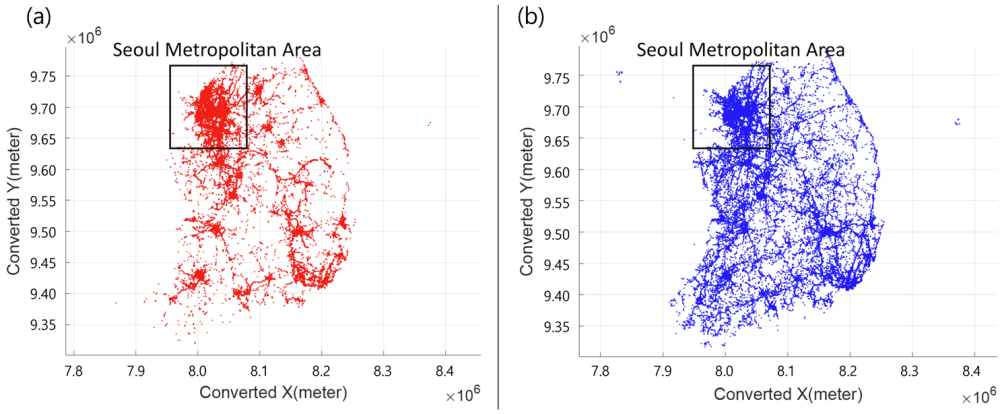

Figure 3. (**a**) Citizen science data, (**b**) fatal accident data.

5. Results

5.1. Results for the Spatial Autocorrelation Analysis

To perform the spatial autocorrelation analysis of inconvenience report counts and fatal accident counts, the study area was partitioned into grid cells. A spatial grid of 50 m × 50 m was applied, which is the smallest unit of geographic information available in Korea. Figure 4a,b illustrates the total sum of inconvenience reports and accident counts in the Seoul metropolitan area visualized for each grid cell after partitioning the entire analysis area into defined grids. Figure 4c,d shows the results after calculating the local Moran's I coefficients for each grid according to Equations (1) and (2).

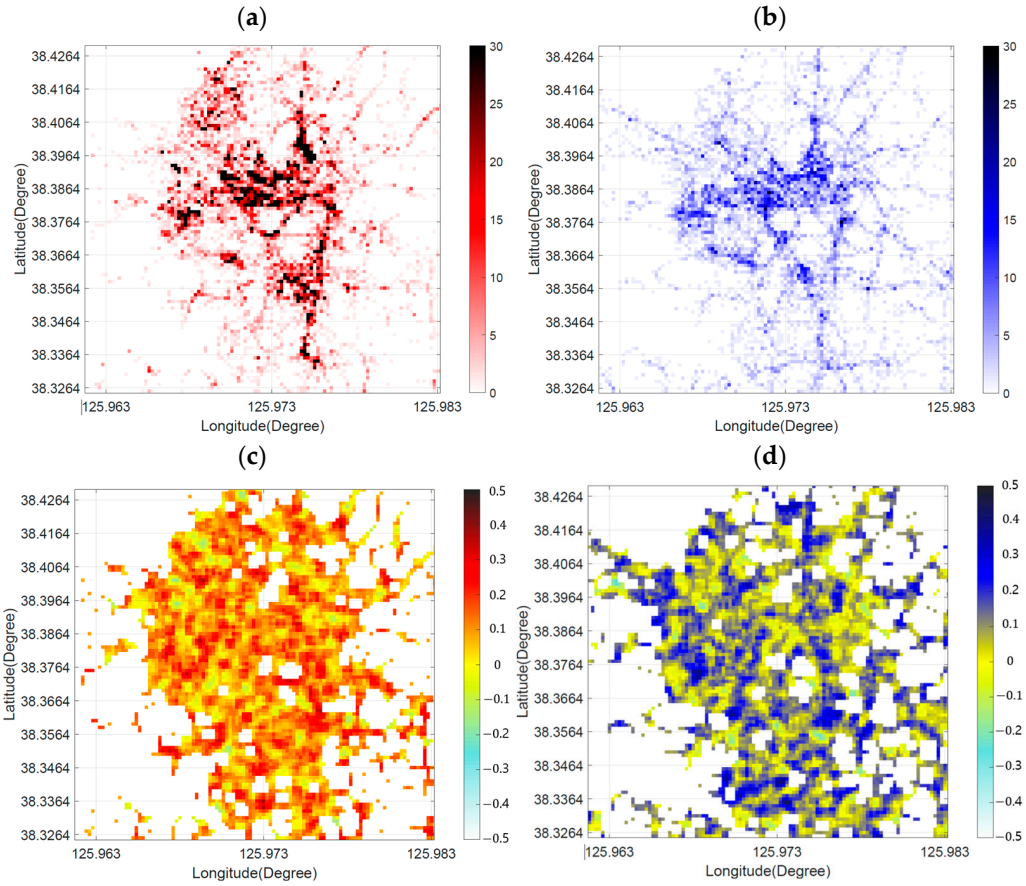

Figure 4. Spatial grid analysis in the Seoul metropolitan area showing (**a**) number of inconvenience reports, (**b**) number of fatal accident occurrences, (**c**) local Moran's I coefficient for inconvenience reports, (**d**) local Moran's I coefficient for fatal accidents.

The local Moran's I coefficient ranges from −1 to 1. Positive values of local Moran's I indicate a cluster of similar values (high or low) around a given location. In Figure 4c, the distribution of local Moran's I for the inconvenience report counts is shown, with positive Moran's I values depicted in red and negative values in cyan. Figure 4d highlights the positive values of local Moran's I in blue. Both figures visually reveal numerous areas with positive local Moran's I values, suggesting that grids with similar attribute values are spatially adjacent. To validate these visual analysis results statistically, this study conducted a right-tailed *t*-test to evaluate the null hypothesis that the local Moran's I from each dataset is sampled from a population with a mean of 0, against the alternative hypothesis that the population mean exceeds 0. The results are presented in Table 3.

Table 3. Statistical analysis results of local Moran's I for citizen science data and fatal accident data.

	Mean	Std	t	*p*-Value ($p < 0.05$)
Citizen Science Data	0.1058	0.0822	163.15	0.000
Fatal Accident Data	0.1022	0.0882	203.98	0.000

According to Table 3, both citizen science data and fatal accident data rejected the null hypothesis that "the local Moran's I value is 0", confirming that the mean values were positive. This indicates that there is spatial autocorrelation in the distribution of both inconvenience reports and fatal accidents. In other words, the distribution of inconvenience reports and fatal accidents is not random; rather, high frequencies in specific areas suggest that adjacent spatial areas are also influenced by these frequencies.

5.2. Results for the Hotspot Analysis

After examining the spatial characteristics of the inconvenience record and fatal accident datasets, the density-based clustering technique DBSCAN was applied to perform hotspot analysis on each dataset. Two parameters must be determined for this technique: epsilon and minimum points. To extract as many hotspots as possible, the minimum points are set to 2, and epsilon is set to 243 m. As a result, a total of 4449 hotspots for inconvenience records and 3077 for fatal accidents are identified. Figure 5a shows the hotspot analysis results from inconvenience records collected in one city in South Korea. Figure 5b displays the results from the fatal accident dataset in the same area. Figure 5c combines the results from Figure 5a,b, revealing areas where the hotspots from both datasets overlap.

In this study, a hotspot shown in Figure 5a is referred to as HS_{inc}, while a hotspot in Figure 5b is referred to as HS_{acc}. In Figure 5c, each HS_{inc} and HS_{acc} can be categorized into two groups depending on whether overlapping areas exist. Thus, an HS_{inc} (or HS_{acc}) that overlaps with an HS_{acc} (or HS_{inc}) is referred to as $HS_{inc,acc}$ (or $HS_{acc,inc}$). Table 4 provides a summary of the number of hotspots from the two datasets and the areas occupied by those hotspots.

Table 4. Number of hotspots by region and data.

Region	$HS_{inc,acc}$	HS_{inc}	$HS_{inc,acc} \times 100/HS_{inc}$	$HS_{acc,inc}$	HS_{acc}	$HS_{acc,inc} \times 100/HS_{acc}$
1	32	328	10	27	96	28
2	163	1299	13	128	664	19
3	64	405	16	47	222	21
4	36	256	14	29	259	11
5	54	185	29	45	88	51
6	39	189	21	35	161	22
7	35	160	22	31	97	32
8	14	68	21	12	172	7
9	205	739	28	166	360	46
10	0	32	0	0	8	0
11	16	66	24	15	72	21
12	30	188	16	30	128	23
13	15	126	12	16	180	9
14	16	135	12	9	174	5
15	2	13	15	1	49	2
16	20	136	15	22	214	10
17	17	124	14	17	133	13
Total	758	4449	17	630	3077	20

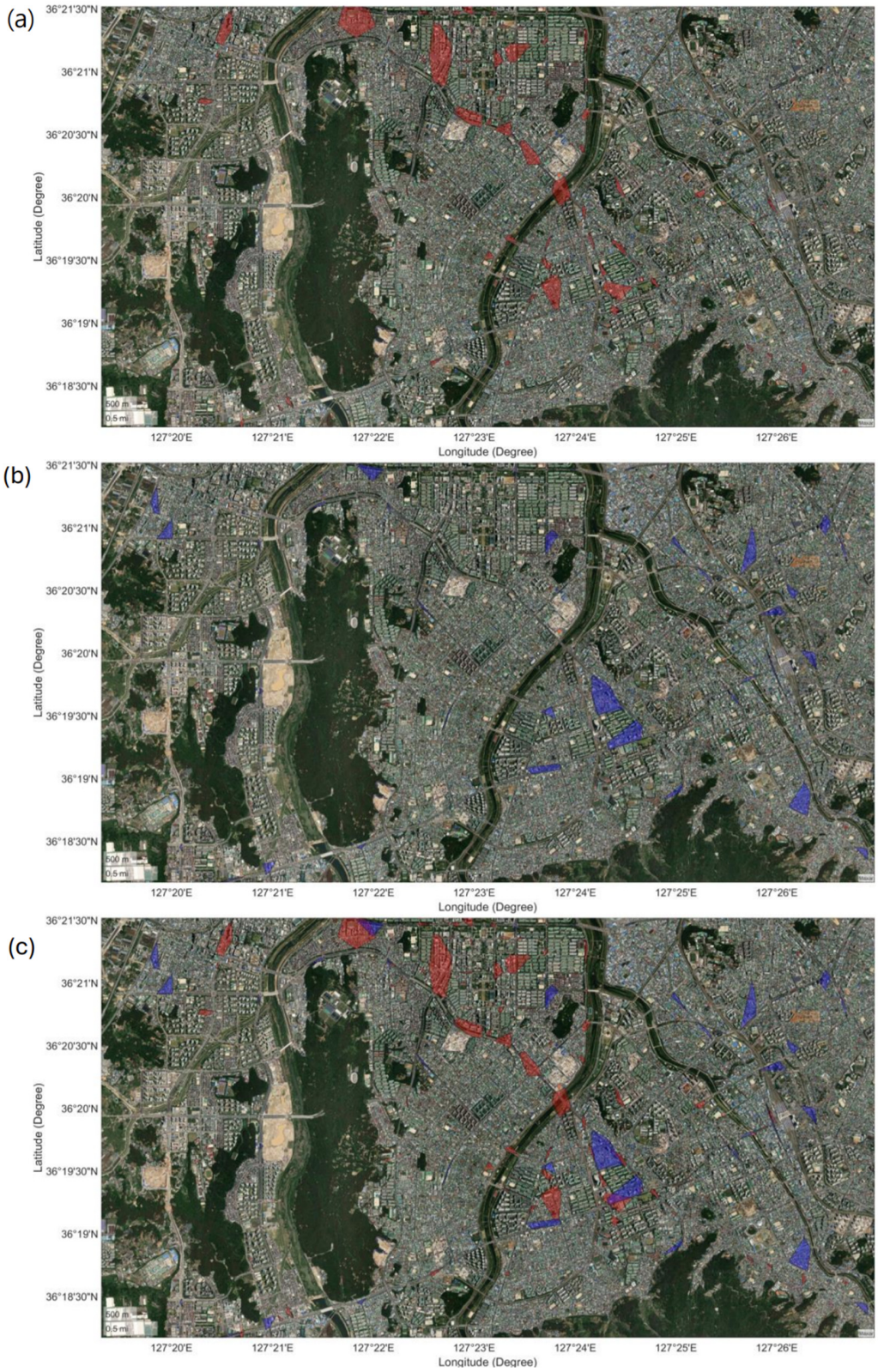

Figure 5. (**a**) Complaint data-based hotspots, (**b**) fatal accident data-based hotspots, (**c**) Complaint + fatal accident data-based hotspots.

When investigating HS_{inc} and $HS_{inc,acc}$, a total of 4449 hotspots were identified as HS_{inc}, of which 758 were classified as $HS_{inc,acc}$. Approximately 17% of the hotspots identified from the inconvenience reports overlapped with the hotspots derived from the accident

data. Conversely, a total of 3077 hotspots were identified as HS_{acc}, with 630 classified as $HS_{acc,inc}$. Around 20% of the hotspots from the accident data overlapped with those from the inconvenience reports. These findings suggest that areas with a high concentration of inconvenience reports do not necessarily correspond to areas with frequent accidents. Nevertheless, 20% of the HS_{inc} remain adjacent to HS_{acc}. If these overlapping hotspots ($HS_{inc,acc}$) exhibit distinct characteristics that differentiate them from other HS_{inc}, identifying these features could provide important insights for predicting high-risk road sections.

To identify features that distinguish $HS_{inc,acc}$ from HS_{inc}, the types of inconvenience reports were examined. As indicated in Table 1, the inconvenience reports were classified into six categories. Among these, 'poor road surface condition', 'potholes', 'roadkill', and 'falling rocks' are considered more likely to induce accidents compared to the other three types. Table 5 presents the results of an investigation into the share of inconvenience reports closely associated with accidents in $HS_{inc,\ acc}$ and HS_{acc} across 17 regions.

Table 5. Number of inconvenience reports within $HS_{inc,\ acc}$ and HS_{inc} by region and type.

Region	Number of Reports in $HS_{inc,acc}$		% of Acc. Related	Number of Reports in HS_{inc}		% of Acc. Related	Diff.**
	Acc. Related*	Others		Acc. Related	Others		
1	38	43	47	1157	996	54	−7%
2	270	195	58	4682	3453	58	+1%
3	88	104	46	1047	1374	43	+3%
4	132	92	59	883	715	55	+4%
5	145	84	63	628	413	60	+3%
6	75	61	55	617	412	60	−5%
7	70	114	38	710	1429	33	+5%
8	14	15	48	165	121	58	−9%
9	750	728	51	1793	2913	38	+13%
10	0	0	0	85	78	52	-
11	47	19	71	137	137	50	+21%
12	54	23	70	590	658	47	+23%
13	18	20	47	316	652	33	+15%
14	47	14	77	352	317	53	+24%
15	5	2	71	25	19	57	+15%
16	35	23	60	365	367	50	+10%
17	43	31	58	400	373	52	+6%
Total	1831	1568	54	13,952	14,427	49	+5%

Acc. related*: Inconvenience reports closely related to accidents. Diff.**: Difference in % between accident related reports in $HS_{inc,acc}$ and in $HS_{inc.}$

According to Table 5, reports associated with accidents constituted 54% of the reports in $HS_{inc,\ acc}$, compared to 49% in HS_{inc}, reflecting a 5% difference. Similar results were observed in 14 out of the 17 regions. This finding suggests that more inconvenience reports associated with accidents are present in areas where actual accidents occurred. These findings can inform policy decisions aimed at accident prevention. For instance, in hotspots where reports of 'poor road surface condition', 'potholes', 'roadkill', and 'falling rocks' comprise a significant portion, addressing these complaints promptly could help prevent accidents.

5.3. Results from the Density Analysis

The results derived from the hotspot analysis alone did not clearly establish the relationship between the frequency of inconvenience reports and fatal accident occurrences. Therefore, this study focuses on $HS_{inc,acc}$ and investigates the density of inconvenience reports. An $HS_{inc,acc}$ includes an area that intersects with $HS_{acc,inc}$, as shown in Figure 6a.

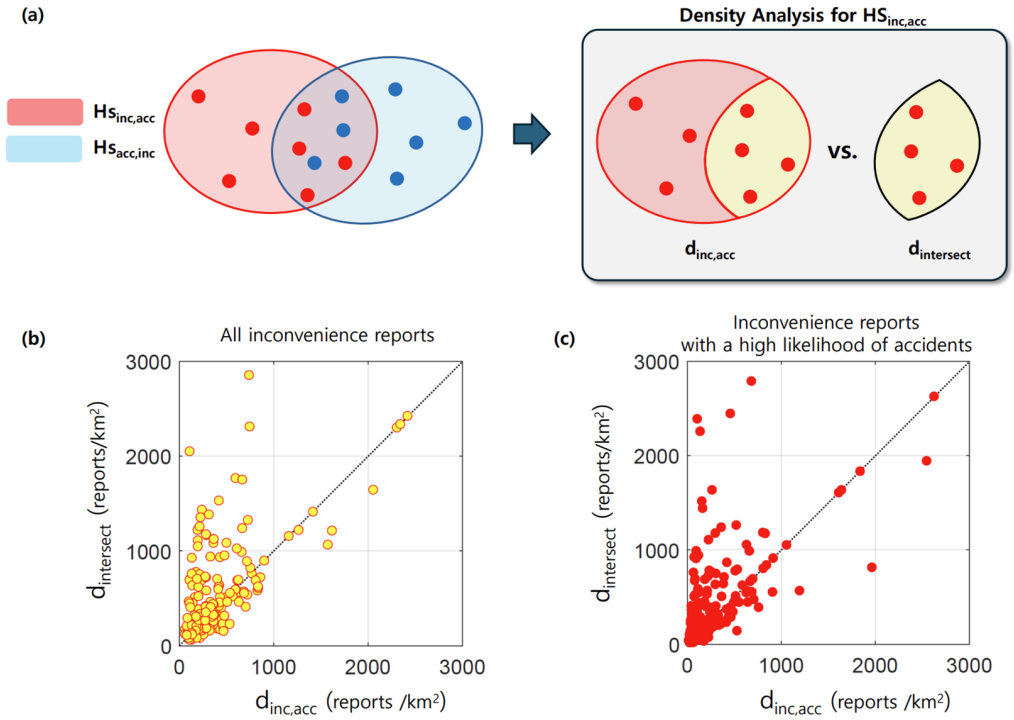

Figure 6. Density analysis results. (**a**) Areas for density analysis, (**b**) Comparison between $d_{intersect}$ and $d_{inc,acc}$ for all inconvenience reports, (**c**) Comparison between $d_{intersect}$ and $d_{inc,acc}$ for a high likelihood of accidents.

In Figure 6a, if the inconvenience reports within $HS_{inc,acc}$ are uniformly distributed spatially, then $d_{inc,acc}$ and $d_{intersect}$ in Figure 6 should be the same. However, if $d_{intersect}$ is greater than $d_{inc,acc}$, this suggests that inconvenience reports are not evenly distributed and tend to cluster near areas with frequent incidents, $HS_{acc,inc}$.

Figure 6b shows the comparison between $d_{intersect}$, and $d_{inc,acc}$ for all types of inconvenience reports, while Figure 6c focuses on report types with a high likelihood of accidents. In both figures, most points fall to the left of the diagonal line, indicating that the densities from the intersection are higher than the densities in $HS_{inc,acc}$. Specifically, the total number of samples on the left side of the diagonal line is 66% in Figure 6b and 66% in Figure 6c. This suggests that inconvenience reports occurring near high-risk accident areas can potentially lead to actual accidents, indicating that special attention should be given to addressing user feedback.

6. Conclusions

Identifying a single cause of traffic accidents is challenging. While some accidents can be attributed to isolated factors such as drunk driving or speeding, most result from a complex interplay of various elements. To address this issue, multiple data sources, including traffic conditions (e.g., traffic flow, speed), road characteristics (e.g., road type, number of lanes), and environmental data (e.g., weather conditions), are integrated to identify locations with a high crash incidence. Additionally, inconvenience reports from road users can serve as valuable indicators of potential accidents. Locations with frequent user reports or a high density of such reports require close inspection and maintenance by road administrators to mitigate potential accident-related factors.

As the proliferation of mobile phones continues, crowdsourcing data have emerged as a promising alternative for addressing traffic accident issues by leveraging public observations and insights. Above all, crowdsourced data have a significant advantage in that they provides sustainable, scalable, and cost-effective insights for traffic safety. This study aims to test the potential of crowdsourcing data to identify segments or areas with a high frequency of accidents. By integrating crowdsourcing data with accident records, we

investigate whether traffic accidents are more likely to occur in areas with a high density of user inconvenience reports.

Various spatial analysis techniques were employed to examine the two datasets. Initially, spatial autocorrelation analysis was performed to evaluate the feasibility of cluster analysis, confirming its applicability to both datasets. Density-based clustering methods were then utilized to identify hotspots based on both complaint and fatal accident data. Finally, a comprehensive analysis of the crowdsourcing data and fatal accident data within overlapping hotspot areas revealed several key findings that substantiate the research hypothesis:

- Spatial autocorrelation analysis revealed that inconvenience reports and accident events were spatially clustered separately.
- Hotspots of inconvenience reports near high-risk accident areas exhibited a higher concentration of accident-related reports.
- Inconvenience reports are not uniformly distributed; rather, they tend to cluster near high-risk accident locations.
- Density analysis demonstrated that traffic accidents tend to occur in areas with frequent inconvenience reports.

This study has clear limitations in explaining the underlying causes of the correlation between user feedback and traffic accidents. Whereas more in-depth research is needed to identify the causes, we speculate that the Heinrich's law can help explain the results of this study. Interestingly, recent studies [34,35] have shown that Heinrich's law can also be applied to events such as traffic accidents.

In practice, identifying the causes contributing to high-risk accident areas is challenging. However, if high-risk accident areas are located near hotspots of inconvenience reports, this study suggests that the reported risks are closely related to accidents occurring in these locations, indicating a need for immediate action to mitigate the identified risks. The significance of this study lies not in quantifying the correlation between inconvenience reports and traffic accidents, but in empirically exploring the potential application of crowdsourcing data in relation to traffic safety. This research provides foundational insights into the quantitative relationship between user feedback regarding road inconvenience and accident occurrences.

Nevertheless, several directions for future research are apparent. Firstly, it is essential to gather additional crowdsourcing data to enhance the empirical foundation. Furthermore, developing alternative analytical methodologies is crucial to address issues related to under- or over-clustering identified during the cluster analysis.

Author Contributions: Study conception and design: S.K., J.K. and W.J.; data collection: J.K. and W.J.; analysis and interpretation of results: S.K. and J.K.; draft manuscript preparation: S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Highway Statistics 2022. Available online: https://www.fhwa.dot.gov/policyinformation/statistics/2022/ (accessed on 9 May 2024).
2. Regan, M.A.; Oviedo-Trespalacios, O. Driver distraction: Mechanisms, evidence, prevention, and mitigation. In *The Vision Zero Handbook: Theory, Technology and Management for a Zero Casualty Policy*; Springer: Cham, Switzerland, 2022; pp. 995–1056. [CrossRef]

3. 1481 Deaths, 3103 Injured: Pothole-Related Accidents on the Rise, Says Report. Available online: https://english.mathrubhumi.com/news/kerala/1-481-dead-3-103-injured-union-govt-report-shows-pothole-related-accidentson-increase-1.8187782 (accessed on 11 May 2023).
4. Setyawan, A.; Kusdiantoro, I. The effect of pavement condition on vehicle speeds and motor vehicles emissions. *Procedia Eng.* **2015**, *125*, 424–430. [CrossRef]
5. El Ferouali, S.; Elamrani Abou Elassad, Z.; Abdali, A. Understanding the Factors Contributing to Traffic Accidents: Survey and Taxonomy. In *The International Conference on Artificial Intelligence and Smart Environment*; Springer: Cham, Switzerland, 2023; pp. 214–221.
6. Yang, F.; Jin, P.J.; Cheng, Y.; Zhang, J.; Ran, B. Origin-destination estimation for non-commuting trips using location-based social networking data. *Int. J. Sustain. Transp.* **2015**, *9*, 551–564. [CrossRef]
7. Hasan, S.; Ukkusuri, S.V. Urban activity pattern classification using topic models from online geo-location data. *Transp. Res. Part C Emerg. Technol.* **2014**, *44*, 363–381. [CrossRef]
8. Zhang, Z.; He, Q.; Gao, J.; Ni, M. A deep learning approach for detecting traffic accidents from social media data. *Transp. Res. Part C Emerg. Technol.* **2018**, *86*, 580–596. [CrossRef]
9. Ounoughi, C.; Ben Yahia, S. Data fusion for ITS: A systematic literature review. *Inf. Fusion* **2023**, *89*, 267–291. [CrossRef]
10. Salas, A.; Georgakis, P.; Petalas, Y. Incident detection using data from social media. In Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), Yokohama, Japan, 16–19 October 2017; pp. 751–755. [CrossRef]
11. Pandhare, K.R.; Shah, M.A. Real time road traffic event detection using Twitter and spark. In Proceedings of the 2017 International Conference on Inventive Communication and Computational Technologies, Coimbatore, India, 10–11 March 2017; pp. 445–449. [CrossRef]
12. Lu, H.; Zhu, Y.; Shi, K.; Lv, Y.; Shi, P.; Niu, Z. Using Adverse Weather Data in Social Media to Assist with City-Level Traffic Situation Awareness and Alerting. *Appl. Sci.* **2018**, *8*, 1193. [CrossRef]
13. Dabiri, S.; Heaslip, K. Developing a Twitter-based traffic event detection model using deep learning architectures. *Expert Syst. Appl.* **2019**, *118*, 425–439. [CrossRef]
14. Rettore, P.H.L.; Santos, B.P.; Rigolin, R.; Lopes, F.; Maia, G.; Villas, L.A.; Loureiro, A.A.F. Road Data Enrichment Framework Based on Heterogeneous Data Fusion for ITS. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1751–1766. [CrossRef]
15. Alkouz, B.; Al Aghbari, Z. SNSJam: Road traffic analysis and prediction by fusing data from multiple social networks. *Inf. Process. Manag.* **2020**, *57*, 102139. [CrossRef]
16. Amin-Naseri, M.; Chakraborty, P.; Sharma, A.; Gilbert, S.B.; Hong, M. Evaluating the Reliability, Coverage, and Added Value of Crowdsourced Traffic Incident Reports from Waze. *Transp. Res. Rec.* **2018**, *2672*, 34–43. [CrossRef]
17. Anggraeni, S.R.; Ranggianto, N.A.; Ghozali, I.; Fatichah, C.; Purwitasari, D. Deep Learning Approaches for Multi-Label Incidents Classification from Twitter Textual Information. *J. Inf. Syst. Eng. Bus. Intell.* **2022**, *8*, 31–41. [CrossRef]
18. Gutierrez-Osorio, C.; González, F.A.; Pedraza, C.A. Deep Learning Ensemble Model for the Prediction of Traffic Accidents Using Social Media Data. *Computers* **2022**, *11*, 126. [CrossRef]
19. Neruda, G.A.; Winarko, E. Traffic event detection from Twitter using a combination of CNN and BERT. In Proceedings of the 2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 23–25 October 2021. [CrossRef]
20. Gu, Y.; Khojastehpour, M.; Jia, X.; Han, L.D. Estimating Pavement Condition by Leveraging Crowdsourced Data. *Remote Sens.* **2024**, *16*, 2237. [CrossRef]
21. Liu, Y.; Hoseinzadeh, N.; Gu, Y.; Han, L.D.; Brakewood, C.; Zhang, Z. Evaluating the coverage and spatiotemporal accuracy of crowdsourced reports over time: A case study of Waze event reports in Tennessee. *Transp. Res. Rec.* **2024**, *2678*, 468–481. [CrossRef]
22. Gu, Y.; Liu, Y.; Liu, D.; Han, L.D.; Jia, X. Spatiotemporal kernel density clustering for wide area near Real-Time pothole detection. *Adv. Eng. Inform.* **2024**, *60*, 102351. [CrossRef]
23. Lin, L.; Li, J.; Chen, F.; Ye, J.; Huai, J. Road Traffic Speed Prediction: A Probabilistic Model Fusing Multi-Source Data. *Trans. Knowl. Data Eng.* **2018**, *30*, 1310–1323. [CrossRef]
24. Liao, Y.; Gil, J.; Pereira, R.H.; Yeh, S.; Verendel, V. Disparities in travel times between car and transit: Spatiotemporal patterns in cities. *Sci. Rep.* **2020**, *10*, 4056. [CrossRef]
25. Lin, Y.; Li, R. Real-time traffic accidents post-impact prediction: Based on crowdsourcing data. *Accid. Anal. Prev.* **2020**, *145*, 105696. [CrossRef]
26. Essien, A.; Petrounias, I.; Sampaio, P.; Sampaio, S. A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. *World Wide Web* **2021**, *24*, 1345–1368. [CrossRef]
27. Janež, M.; Verovšek, Š.; Zupančič, T.; Moškon, M. Citizen Science for Traffic Monitoring: Investigating the Potentials for Complementing Traffic Counters with Crowdsourced Data. *Sustainability* **2022**, *14*, 622. [CrossRef]
28. Dienstl, B.; Scholz, J. A Concept for Smart Transportation User-Feedback Utilizing Volunteered Geoinformation Approaches. *Adv. Intell. Syst. Comput.* **2018**, *879*, 538–545. [CrossRef]
29. Liu, Y.; Feng, T. The Effect of Crowdsourced Police Enforcement Data on Traffic Speed: A Case Study of The Netherlands. *Appl. Sci.* **2023**, *13*, 11822. [CrossRef]

30. Klopfenstein, L.C.; Delpriori, S.; Polidori, P.; Sergiacomi, A.; Marcozzi, M.; Boardman, D.; Parfitt, P.; Bogliolo, A. Mobile crowdsensing for road sustainability: Exploitability of publicly-sourced data. *Int. Rev. Appl. Econ.* **2020**, *34*, 650–671. [CrossRef]
31. Alamri, S. The Geospatial Crowd: Emerging Trends and Challenges in Crowdsourced Spatial Analytics. *ISPRS Int. J. Geo Inf.* **2024**, *13*, 168. [CrossRef]
32. Li, X.; Dadashova, B.; Yu, S.; Zhang, Z. Rethinking Highway Safety Analysis by Leveraging Crowdsourced Waze Data. *Sustainability* **2020**, *12*, 10127. [CrossRef]
33. Fotheringham, A.; Rogerson, P. *The SAGE Handbook of Spatial Analysis*; SAGE Publications: Thousand Oaks, CA, USA, 2008. [CrossRef]
34. Cho, S.; Kim, D.; Khan, H.; Kil Lee, C. Heinrich's Law for Traffic Incidents–Using the Digital Tachograph Data to Identify Traffic Accident Hotspots. *Promet Traffic Transportation* **2023**, *35*, 829–837. [CrossRef]
35. Park, J.; Kim, S.; Kim, J. Exploring spatial associations between near-miss and police-reported crashes: The Heinrich's law in traffic safety. *Transp. Res. Interdiscip. Perspect.* **2023**, *19*, 100830. [CrossRef]