How do people perceive driving risks in small towns? A case study in Central Texas

Xiao Li [a,*], Greg Rybarczyk [b,c,d], Wei Li [e], Muhammad Usman [e], Jiahe Bian [f], Andong Chen [e], Xinyue Ye [e]

[a] Transport Studies Unit, University of Oxford, South Parks Road, Oxford OX1 3QY, UK
[b] College of Innovation and Technology, University of Michigan-Flint, Flint, MI 48502, USA
[c] Michigan Institute for Data Science, The University of Michigan, Ann Arbor, MI 48108, USA
[d] The Centre for Urban Design and Mental Health, London SW9 7QF, UK
[e] Department of Landscape Architecture & Urban Planning, Texas A&M University, College Station, TX 77843, USA
[f] School of Planning, University of Cincinnati, Cincinnati, OH 45221, USA

A B S T R A C T

The number of studies investigating the relationship between perceived and objective traffic risk from drivers' perspective is limited. This study aims to investigate this dynamic within an understudied transportation environment – small towns in Texas, USA, defined as incorporated places with a population of less than 50,000. A web-based survey was distributed to six small towns in central Texas to ascertain perceptual traffic risk factors and personal characteristics. A participatory GIS exercise was also conducted to collect where high-risk locations were perceived and to correlate them to high crash zones. This study spatially examined the relations between perceived and observed risk locations and statistically identified a set of contributing factors which could make crash-intensive areas more perceivable by road users. The results indicated that road users' perceived risk locations are not always associated with high crash rates. The match rate between perceived and observed risk locations varied significantly across studied sites. We found that some personal and built environment factors significantly impacted people's sensitivity to perceiving crash-intensive locations. The binary logistic regression model was accurate (74.13%) in highlighting whether a perceived risk location matches observed risk locations. The results emphasize the importance of considering perceived and objective risk simultaneously to gain a better understanding of traffic risk mitigation, especially in underserved small towns.

1. Introduction

Road fatalities continue to be a global public health crisis. According to the World Health Organization, nearly 1.2 million people die each year on roads globally (World Health Organization, 2018), and in low- to middle- income countries, it is the tenth leading cause of death (World Health Organization, 2020). Traffic fatalities disproportionately impact persons aged 15–24 (International Transport Forum, 2018). Efforts to curb these trends is a critical objective, as evidenced by a recent United Nations (UN) road safety strategy (United Nations, 2019). Vision Zero, initiated in Sweden, is focused on reducing traffic deaths in cities by applying a holistic approach to designing safe streets (Ferenchak, 2022). The United States has also recognized this public health crisis and has resultantly helped implement "Complete Streets" programs in many cities, where the goal is to elevate safe travel for all modes, including automobiles (LaPlante and McCann, 2008). Clearly, reducing travel fatalities and lowering risk remains a top priority throughout the world.

Researchers and practitioners have studied traffic risk for decades. Historical crash data (e.g., traffic crash reports) recorded by law enforcement agencies is the primary data source in existing road safety studies. Many studies have been conducted to explore the spatiotemporal distributions (Nie et al., 2015; Rhee et al., 2016), uncover the leading factors (Huang et al., 2018; Song et al., 2006), and forecast the likelihood of objective traffic risks (Cai et al., 2017; Lee et al., 2018) using crash data. However, compared to the intensive studies on objective traffic risks, studies on subjective traffic risks, such as traffic risk perception, are relatively few. Perceived traffic risk is commonly defined as a road user's subjective interpretation of risk when involved

in different traffic situations (Deery, 1999). Researchers have gradually acknowledged the importance of traffic risk perception due to its impacts on many societal levels. For example, traffic risk perception can directly impact road behaviors and consequently impact safety outcomes (Nordfjærn and Rundmo, 2009; Wang et al., 2002). The perception of risk also weighs heavily on neighborhood satisfaction, which is positively correlated with traffic safety (Lee et al., 2017). The connection between physical activity and perceived risk has also received much attention in the literature, albeit with inconclusive results. Therefore, understanding the human mechanisms of how traffic risk is conceived of and realized can have significant implications for road safety planning and enhancement (AlKheder et al., 2022; Kononov et al., 2007; Schneider et al., 2004).

1.1. Related studies

A growing number of studies were conducted to understand the human mechanisms of traffic risk perception, mainly focusing on two topics: (1) the relations with different influential factors (e.g., driving behaviors, built environment, sociodemographic features) and (2) the relations with traffic crashes.

Prior studies have examined the influential factors impacting road users' risk perceptions from different aspects. For example, Machado-León et al. (2016) performed face-to-face interviews with 492 drivers to explore how risky driving behaviors impact drivers' perceptions of crash risks and how the perception varies among socially different drivers. A stated preference ranking survey was performed to measure the respondents' crash risk perceptions. Their results showed that people's risk perception of dangerous driving behaviors is significantly impacted by their driving experience and socioeconomic characteristics, such as income, household size, and gender. AlKheder et al. (2022) conducted a survey to explore what factors impact pedestrians' perceptions and how their perceptions correlate with walking frequencies. Findings showed pedestrians with a higher level of worry in walking have a noticeably lower walking frequency. Time of day and built environment features (e. g., walkability index, the density of shops and businesses) significantly impact pedestrians' perceptions of risk for walking. A similar study by Rankavat and Tiwari (2016) explored how built environment features affect pedestrians' perceptions of crash risks in neighborhoods (where they reside) and crash-intense areas. The results indicated that built environment and traffic features, such as the number of lanes, road median width, sidewalk width, etc., impact perceived travel risk. Although these findings are meaningful for incorporating traffic risk perception into urban/transportation planning, the implications of perceived risk in road safety applications are still unrevealed without linking them with objective traffic risks, such as crashes.

Researchers have made preliminary attempts to compare perceived risk locations and actual crash observations. For example, Karim (1992) invited students and staff members to rank a list of locations selected from a university campus based on their perceived crash risks. The authors found a close similarity between the ranking of perceived risks and actual crash counts for these locations, indicating that high-risk areas are perceivable by road users. In a similar study, Lee et al. (2016) surveyed school-aged children from eight elementary schools, using stickers on printed maps indicating high-risk locations in their school zones. These pinpointed locations were spatially aggregated to the closest road intersections; the counts of pinpointed locations for the intersections were used to represent the perceived risks by students. The authors created different models based on the intersection characteristics (e.g., exposure, road infrastructure, traffic signs) to estimate the number of pinpointed locations and crashes for each intersection. The results showed that some variables remained consistent relationships with both perceived and observed risk measures, implying that the risk perceptions of school-aged children are relatively accurate. Similarly, von Stülpnagel and Lucas (2020) examined cycling crash risk observations and perceptions. They collected crowdsourced hazards reported by

cyclists to represent the perceived cycling risks, which were then compared with police-reported bike crashes. Models were created to estimate the number of crowdsourced hazards and bike crashes for intersections. This study found a high similarity of contributing factors for objective and subjective cycling risks. However, Schneider et al. (2004) reported different results. They examined the distribution of police-reported pedestrian crashes and locations with high perceived risk within a university campus. A total of 312 pedestrians and 110 drivers were invited to mark risky locations on printed maps through a mail survey. They found the distributions of actual crashes and perceived risky locations were significantly different and attributed to varying levels of traffic exposure, crosswalk density, and sidewalk conditions.

1.2. Knowledge gaps and research objectives

Although some efforts have been made in the past to compare the relationship between objective traffic risks (e.g., perceived risk locations) and subjective traffic risks (e.g., crashes), several important questions remain unanswered. First, most studies used crash count/ frequency to represent objective traffic risks and compared them with perceived traffic risks. Note that many studies have demonstrated that using crash count/frequency to quantify traffic risk is problematic due to the fact that it does not take into account the amount of traffic flow, potentially misclassifying busy (high-volume) but safe segments/areas as high risk (Li et al., 2021a; Yao et al., 2016). Therefore, comparing perceived traffic risks with appropriate objective traffic risk measures requires additional exploration. Second, studies have shown that perceived risk locations may not match up with the observed dangerous locations in terms of crash risks (Duncan and Hughes, 2002), which poses a new question: What factors contribute to the "spatial match" between perceived traffic risk and observed crash risk? Third, most prior studies were conducted in highly populated areas (e.g., urban areas or university campuses). Many studies have indicated that rural and underserved communities are more vulnerable and thus more likely to observe crashes (Chimba et al., 2018; Li et al., 2022). This could be attributed to differences in road design (Blatt and Furman, 1998), built environment (Cabrera-Arnau et al., 2020; Svenson et al., 1996), as well as drivers' driving attitude and risk perception (Rakauskas et al., 2009) between rural and urban areas. However, to date, there is a dearth of investigations on perceived traffic risk in underserved communities, which requires more attention.

These unanswered questions leave us with an incomplete understanding of the (geographical) relations between perceived and objective traffic risks and the influential factors contributing to the "spatial match" between them, especially in small and rural communities. While we build on earlier studies of this genre (Noland, 1995; Rankavat and Tiwari, 2020; Schneider et al., 2004), this study set out to collectively address the aforementioned research gaps. Thus, this research has three objectives: (a) descriptively describe perceived risk from persons residing in six rural Texas small towns; (b) empirically and spatially explore perceived and observed associations and patterns; and (c) estimate the binary spatial relationship (matched or unmatched) between perceived and observed traffic risk locations using a logistic regression model accounting for perceived, personal/household, and neighborhood factors.

2. Study area and data

2.1. Study sites-Six small towns in Central Texas

The study areas include six small towns in Central Texas, including Caldwell, Copperas Cove, Harker Heights, Huntsville, Madisonville, and Nolanville, as illustrated in Fig. 1. Note that this study defines small towns as incorporated places with<50,000 population (The United States Office of Management and Budget, 2022); we also included rural villages that share the same postal distribution routes. We chose these
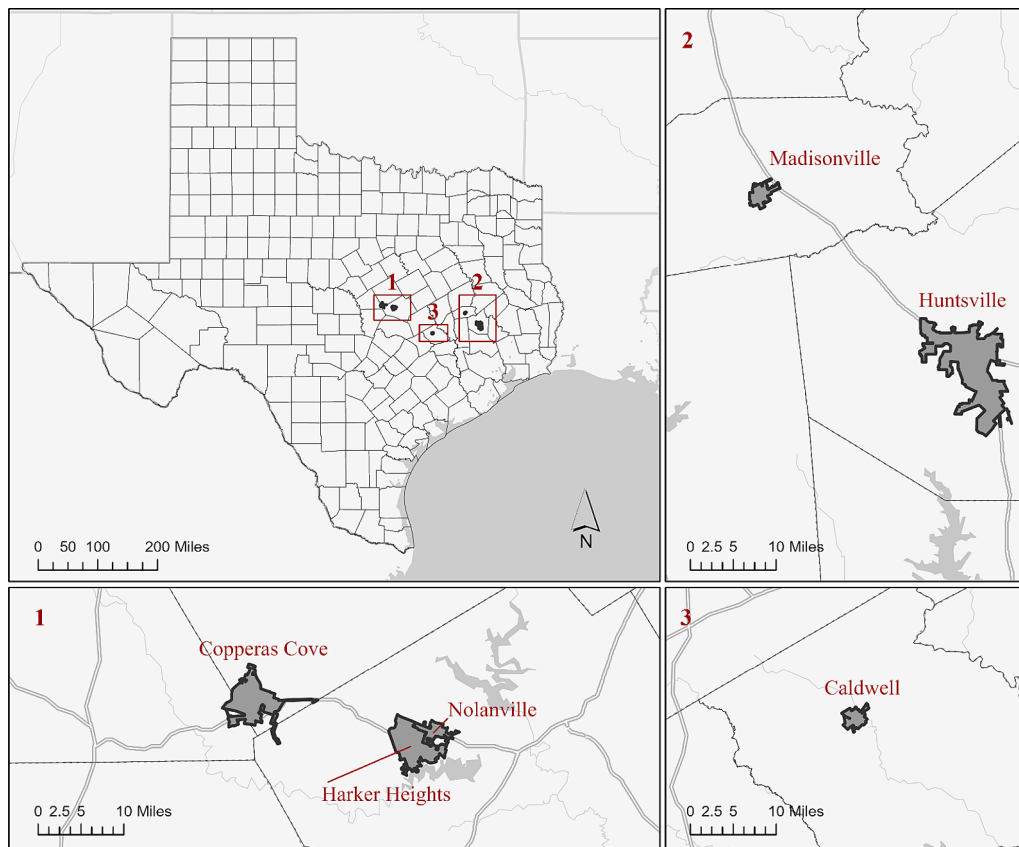
Fig. 1. Selected rural towns in Texas, USA.

sites because they are typically small and rural communities with population sizes ranging from 3,993 to 45,941 (United States Census Bureau, 2021). Our sampled towns contain a relatively high percentage of low-income; their demographic and economic developments are projected to grow steadily compared to the Texas average (Austin Capital Advisors, 2021).

2.2. Data sources

This research integrated four types of data to match previous studies, including survey-based perception data, historical crash records, road inventory data depicting the geometric design of roadways, and the neighborhood features characterizing built environment and location efficiency (i.e., density of development, diversity of land use, accessibility to destinations) relevant information (Table 1).

The web-based survey collected respondents' perceived risk

Table 1
Data sources and purpose.

Data Type	Data Source	Purpose
Perception Data	Web-based survey (August 2021 - May 2022)	Collecting perceived risk locations, perceived risk factors, and survey respondents' personal information.
Crash Records	TxDOT CRIS (2016–2020)	Calculating crash rate and characterizing crash relevant information
Road Inventory	TxDOT RHiNo (2020)	Calculating crash rate and characterizing roadway relevant information
Built Environment and Location Efficiency	U.S. EPA SLD (2021)	Characterizing built environment and location efficiency relevant information

locations, risk factors, and their personal information from the six selected small towns, detailed in Section 3.1. To quantify the observed crash risk, zonal crash rates were calculated based on officially recorded crash records and traffic exposure data, such as road length and traffic volume. This study retrieved five years (2016–2020) of crashes from the Texas Department of Transportation (TxDOT) Crash Records Information System (CRIS). To control for neighborhood effects on the spatial relation between perceived and observed risk locations, we collected and generated a list of explanatory variables from different government sources. The roadway relevant-explanatory variables were generated based on the roadway data from TxDOT RHiNo 2020. Built environment & location efficiency variables were retrieved from the Smart Location Database 2021 (SLD 2021) released by the United States Environmental Protection Agency (U.S. EPA). The U.S. EPA SLD contains 90 different socioeconomic variables and environmental correlates aggregated to the Census Block Group (CBG) level for the entire nation (Ramsey and Bell, 2014). This database has been used in similar studies on travel behavior and planning analysis (Rybarczyk et al., 2018; Yang et al., 2020). We want to emphasize that while the data collection periods for these datasets are not perfectly synchronous, we are confident that the crash data (2016–2020), roadway characteristics data (2020), and built environment data (2021) offer reliable insights into safety outcomes, road infrastructures, and neighborhood attributes that are unlikely to have undergone significant changes over a few years. The survey data collected from August 2021 to May 2022 can effectively capture their impacts on people's perceptions. The impact of the asynchronous timing is this study is minimal.

3. Methods

This study was conducted in four main steps, as illustrated in Fig. 2. First, we conducted a web-based survey to collect local residents'
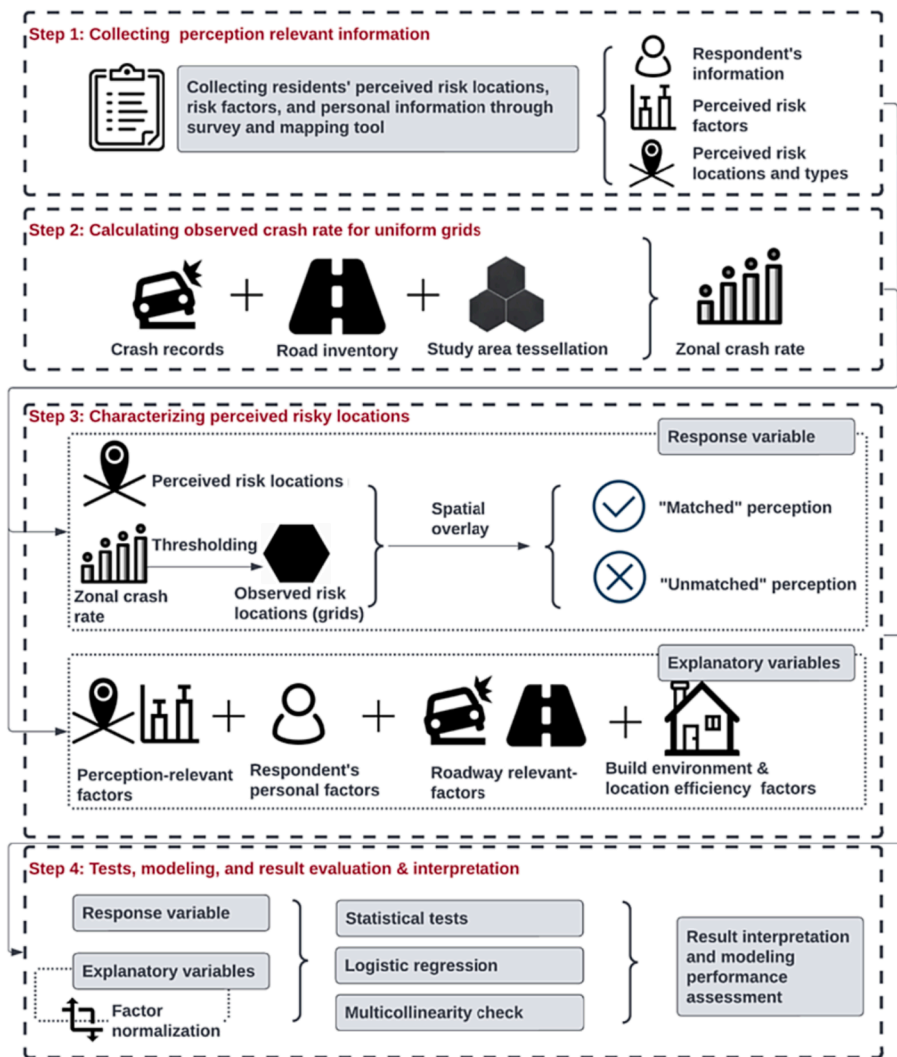
Fig. 2. Research workflow.

perceived risk locations, perceived risk factors, and personal information. Next, crash and road inventory data were used to calculate the crash rates for tessellated uniform grids, which were used as the objective risk observations to compare with the perceived risk locations. Then, we labeled each perceived risk location as "matched" or "unmatched" based on whether it spatially overlaps with any high crash risk grids determined by the crash rates. Meanwhile, we collected and generated a list of features to characterize each perceived risk location from four perspectives: perception-relevant factors, respondent's personal factors, roadway-relevant factors, and built environment & location efficiency factors (see details in Section 3.3.2.). Last, we performed statistical tests to assess what factors significantly impact the binary spatial relations ("matched" and "unmatched") between perceived risk locations and observed risk grids, then applied logistic regression to model this relationship. Different approaches were employed to reduce irrelevant and intercorrelated explanatory variables. The following sections highlight these steps in detail.

3.1. Web-based survey for perception data collection

We constructed a web-based survey to collect respondents' demographic details, perceived risk locations, and the corresponding perceived risk factors associated with each location. Since one of the main purposes of the survey was to ask respondents to geolocate the locations with traffic safety concerns, a map-based response was

desirable as it could provide all the necessary geographical details. Therefore, we opted for a web-mapping tool—Maptionnaire, which allows participants to map their perceived risk locations and further specify the types of location (e.g., a place of interest [POI], an intersection, a road, or a neighborhood) and the perceived risk factors through associated questions (Fig. 3). The survey was filled out completely using Maptionnaire and then linked with Qualtrics. Qualtrics was used to ascertain each respondent's personally identifiable information, such as demographic information, the number of operable household automobiles, valid driver's license, the time elapsed since acquiring a license, any traffic citations in the past two years, age, gender, marital status, household income, employment status, and resident's current city. Both the databases from Maptionnaire and Qualtrics were then exported and joined to create one combined database of all the information. Through Maptionnaire, participants reported up to three high-risk locations based on their perception and answered associated questions for each location to specify the location type and the accompanying reasons, such as poor road surface, high speed limit, aggressive drivers, heavy traffic, poor lighting, poor quality surrounding environment, or too many pedestrians and/or bicyclists. Survey records with low quality responses (e.g., extremely short response time, many unanswered questions, pinpointed locations outside of their residential areas) were excluded from this study.

In this study, an IRB approval for human subject research was obtained from the institutions with which the authors are affiliated. To be
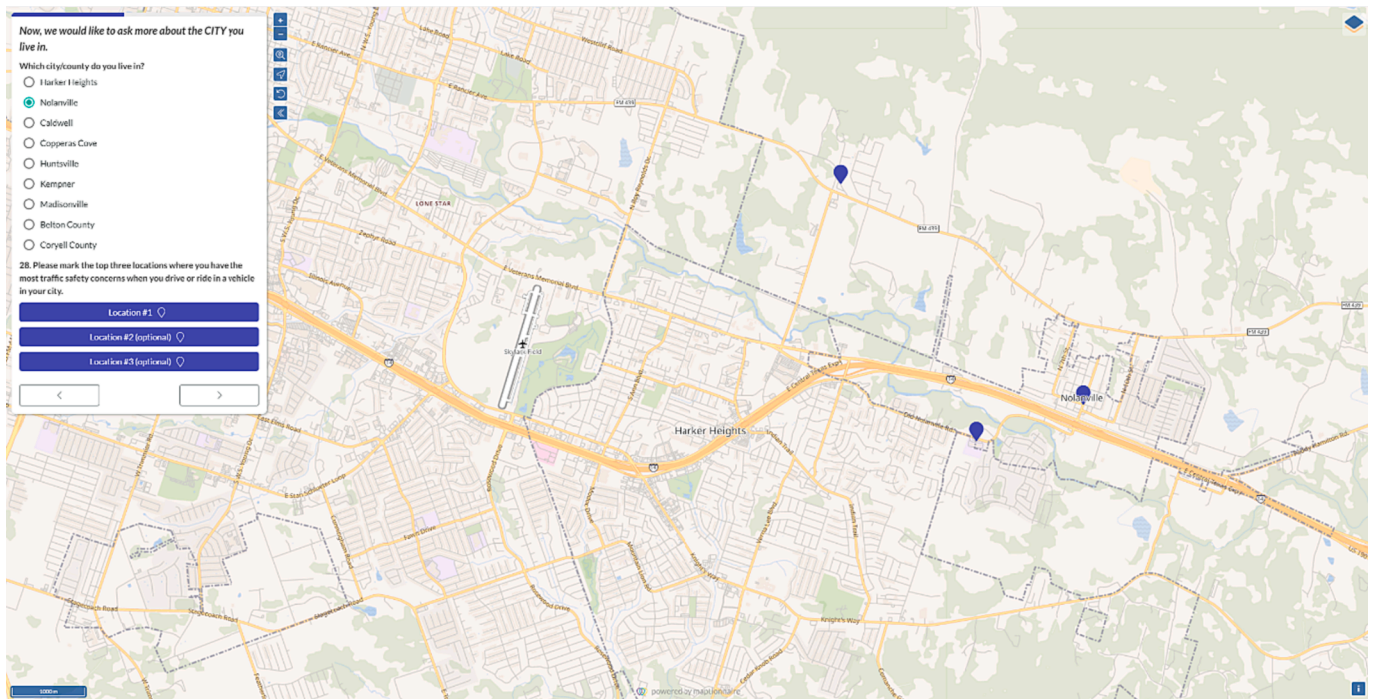
Fig. 3. Example of Maptionnaire online survey for collecting perceived risk locations.

eligible for the survey, participants needed to be 18 years of age or older and residents of the selected six cities, possessing an inherent understanding of the city's traffic context. Only one participant is allowed per household. Survey flyers were distributed using the United States Postal Service Every Door Direct Mail (EDDM) service within the study sites, which contain the link to the web-mapping tool—Maptionnaire to pinpoint perceived risk locations.

3.2. Zonal equivalent property damage only (EPDO) crash rate derivation

As introduced above, although respondents' perceived risk locations were mapped as points, they could represent four types of location with different vector types (e.g., point: POI, intersection; line: road; polygon: neighborhood). Meanwhile, since these locations were manually pinpointed, positioning errors could be introduced, making it difficult to pinpoint them precisely on the map, especially for the point and line locations. To effectively accommodate the positioning errors and compare with different types of locations, we divided each city into equal-sized tessellated grids and calculated the Equivalent Property Damage Only (EPDO) crash rate for each grid to represent the observed crash risk.

Two factors need to be considered during the tessellation process: grid shape and size. This study chose to use the hexagonal grid during the tessellation process. Many studies examined the effects of different geometric zoning systems on transportation studies (Chmielewski and Kempa, 2020; Ghadiri et al., 2019). Compared to other commonly used grid shapes (e.g., square, rhombus, triangle), the hexagon is most similar to a circle and outperforms other shaped grids in transportation studies, such as planning (Chmielewski and Kempa, 2020) and trip production models (Ghadiri et al., 2019). Concerning grid size, since each mapped point could represent a location ranging from a POI to a neighborhood, we believe census blocks—the smallest census data unit—could be comparable to perceived risk locations. Please note that the size of census blocks in urban areas is much smaller than in rural areas. Given that most perceived locations were collected from small Texas cities, generally classified as urbanized areas, we used the average size of census blocks in Texas urban areas statewide (0.05 square mile) as the grid size.

A crucial step in traffic safety studies is to select and generate appropriate safety performance measures. The crash rate is one of the most used safety measures, which quantifies crash risk by normalizing the crash counts based on traffic exposure (The U.S. National Highway Safety Administration, 2019). According to the definition provided by the Federal Highway Administration (FHWA), the crash rate for roadway represents the number of crashes for every 100 million vehicle miles traveled (VMT), which can be calculated through Equation (1) (The U.S. National Highway Safety Administration, 2019):

$$R_{seg} = \frac{C_{seg} * 100,000,000}{365 * N * V * L} \tag{1}$$

where R_{seg} represents the segment-level crash rate; C_{seg} is the count of crashes occurred on the segment; N represents the number of years; V indicates the traffic volume (AADT) of the segment; L indicates the segment length (in miles).

One limitation associated with the traditional crash rate calculation is that it fails to account for crash severity. In response to this issue, the EPDO method was proposed, aiming to incorporate crash severity by assigning weights to different types of crashes (e.g., fatal, injury, and property damage-only [PDO]) based on their societal costs. This results in the development of an EPDO score. As recommended by the 2010 Highway Safety Manual, the societal cost of a fatal crash is estimated to be $4,008,900, equivalent to 541.7 PDO crashes in cost; the societal cost of an injury crash is $82,600, equivalent to 11.2 PDO crashes (American Association of State Highway and Transportation Officials, 2010). Therefore, the EPDO score (*EPDO*) for a specific site can be expressed using Equation (2):

$$EPDO = 541.7 * Fatal\ Crashes + 11.2 * Injury\ Crashes + PDO\ Crashes \tag{2}$$

In this study, we integrated the EPDO method with crash rate computation to derive the zonal EPDO crash rate—normalizing the EPDO score by considering the aggregated traffic exposure within each gird using Equations (3–4) (Li et al., 2022):

$$R_{grid} = \frac{EPDO_{grid} * 100,000,000}{E_{grid}} \tag{3}$$

$$E_{grid} = \sum_{j=1}^{r} \sum_{i=1}^{n} L_i * V_{ij} * 365 \qquad (4)$$

where R_{grid} indicates the zonal EPDO crash rate of a hexagonal gird; $EPDO_{grid}$ is the EPDO score for the grid during the study period; E_{grid} is the aggregated grid-level traffic exposure; n represents the number of segments within the grid; i is the i-th segment; r represents time span of the crash data (in years), set as 5 in this study; j is the j-th year, L_i is the length of i-th segment (in miles), and V_{ij} indicates the traffic volume of the i-th segment at the j-th year. In accordance with past research, we selected the top 20% of grids in the ranking of their crash rates to represent the observed risk locations in each of the study cities (Li et al., 2021a; Thakali et al., 2015; Yu et al., 2014).

3.3. Classification and characterization of perceived risk locations

3.3.1. Generating the binary dependent variable

After obtaining the observed risk locations through the crash rate calculation, we performed the spatial overlay analysis on the perceived risk locations (points) and observed risk locations (grids) to check whether perceived risk points fell in or outside of the high-risk hexagonal grids. The result was labeled "matched" (1) or "unmatched" (0). This binary result was considered the dependent variable in our regression model.

3.3.2. Characterization of perceived risk locations

Past research has demonstrated that people's perception to traffic risks can be impacted by personal reasons (DeJoy, 1989), character traits (e.g., gender, education, income) (Deery, 1999), road and traffic conditions (Leviäkangas, 1998), and built environment conditions (Ewing and Dumbaugh, 2009). To match this body of work and comprehensively explore the potential factors impacting the "spatial match" between perceived and observed risk locations, we compiled a list of variables to characterize each collected perception records from four perspectives: individual perception, personal attributes, roadway factors, and neighborhood context and composition features. Table 2 lists the initial explanatory variables and their data sources considered in this research.

3.4. Statistical tests, model development, and result evaluation

3.4.1. Exploratory data analysis [EDA]

To meet our first two objectives (a, b), we employed several EDA techniques to understand statistical and geographical traffic risk trends. Traditional descriptive analysis was first applied to participants' perceived risks as they related to location and environmental characteristics. We also created geo-visualizations to showcase spatial relationships between observed risk zones (i.e., hexagonal grids) and perceived traffic risk locales. In addition, chi-squared tests of independence and unpaired two-sample Wilcoxon tests (aka, Wilcoxon rank sum test or Mann-Whitney test) were also applied to assess whether each categorical and numerical explanatory variable was associated with the binary response variable (i.e., "matched" and "unmatched"). The chi-squared test of independence is commonly used to assess whether two categorical variables are likely to be related. The Wilcoxon test is designed to assess differences between two independent groups when the dependent variables are continuous but not normally distributed.

3.4.2. Model development and feature selection

To meet our third objective (c), we developed a binary logistic regression model to investigate the impacts of various variables from four key categories on the probability of a traffic risk match between perceived risk and objectively measured high risk traffic zones. Logistic regression is a well-established statistical method commonly used to estimate the probability of an event occurring based on selected explanatory variables, particularly for modeling binary outcomes

Table 2
List of selected explanatory variables.

Variable Name	Data Type	Value (Categories/ Description)	Data Source
Perception factors			
Perceived location type	Categorical	Place of interest (POI); Intersection; Road; Neighborhood	Survey
Perceived poor road surface	Binary	Yes or No	Survey
Perceived high speed limit	Binary	Yes or No	Survey
Perceived many aggressive drivers	Binary	Yes or No	Survey
Perceived heavy traffic volume	Binary	Yes or No	Survey
Perceived poor lighting	Binary	Yes or No	Survey
Perceived poor surrounding environment	Binary	Yes or No	Survey
Perceived many bikes and pedestrians	Binary	Yes or No	Survey
Personal attributes			
Number of cars within your household	Numerical	Number of vehicles in the respondent's household	Survey
With a valid driving license	Binary	Yes or No	Survey
Length since licensed	Categorical	<1 month; 1 month to 1 year; 1 year to five years; Over five years	Survey
Involved in crash in past two years	Binary	Yes or No	Survey
Received any citation in past two years	Binary	Yes or No	Survey
Age	Categorical	Generation Z (19–25 years old); Millennials (26–41 years old); Generation X (42–57 years old); Boomers (58 +); Unwilling to response	Survey
Gender	Categorical	Female or Male	Survey
Marriage	Categorical	Married or with a partner; Divorced, windowed or separated; Never married	Survey
Education	Categorical	No Bachelor's degree; Bachelor's degree or higher; Unwilling to response	Survey
Total household income	Categorical	Under $50 k; $50 k-$100 k; Over $100 k; Unwilling to response	Survey
Employment	Binary	Yes or No	Survey
Retired	Binary	Yes or No	
City	Categorical	Caldwell; Copperas Cove; Harker Heights; Huntsville; Madisonville; Nolanville	Survey
Roadway factors (aggregated to the gird-level)			
Road density (mile/ square mile)	Numerical	Length of roadway (in mile) per square mile	RHiNo 2020
Traffic exposure	Numerical	The sum of roadways' length multiplies their traffic volume for each grid, obtained using equation (3)	RHiNo 2020
Built environment and location efficiency factors (aggregated at the Census Block Group [CBG] level)			
Percent of zero-car households	Numerical	Percent of zero-car households at the CBG level	SLD 2021
Percent of one-car households	Numerical	Percent of households with one automobile at the CBG level	SLD 2021
Percent of two-plus-car households	Numerical	Percent of households with two or more automobiles at the CBG level	SLD 2021

Table 2 (*continued*)

Variable Name	Data Type	Value (Categories/Description)	Data Source
Gross residential density	Numerical	Gross residential density (household/acre) on unprotected land	SLD 2021
Gross population density	Numerical	Gross population density (people/acre) on unprotected land	SLD 2021
Gross employment density	Numerical	Gross employment density (jobs/acre) on unprotected land	SLD 2021
Jobs per household	Numerical	Jobs per household at the CBG level	SLD 2021
Employment and household entropy	Numerical	The mix of employment types and occupied housing.	SLD 2021
Employment and household entropy (trip based)	Numerical	Employment and Household entropy based on vehicle trip production and trip attractions	SLD 2021
Street intersection density	Numerical	The number of intersections per square mile at the CBG level	SLD 2021
Regional diversity	Numerical	A measure quantifies the deviation of the CBG ratio of jobs/pop from the regional average ratio of jobs/pop	SLD 2021
Jobs within 45 min auto travel time	Numerical	The number of accessible job opportunities within 45 min auto travel time	SLD 2021
Working age population within 45 auto travel time	Numerical	The number of working age population within 45 min auto travel time	SLD 2021
National walkability index	Numerical	A measure quantifying the likelihood and feasibility of walking at the CBG level	SLD 2021
Percent of low-wage workers	Numerical	Percent of low-wage workers at the CBG level	SLD 2021
Percent of medium-wage workers	Numerical	Percent of medium-wage workers at the CBG level	SLD 2021
Percent of high-wage workers	Numerical	Percent of high-wage workers at the CBG level	SLD 2021
Smart location index	Numerical	An index indicating the location efficiency, ranging from 0 to 100. The higher the score, the more efficient the site is.	SLD 2021

(Subasi, 2020). This approach aligns perfectly with the modeling purpose of our study.

To enhance the stability and performance of the logistic regression model, we standardized numerical variables measured at different scales to a range of 0 to 1. This normalization process ensures that all variables contribute equally to the model's performance, thereby reducing the potential for bias and improving the model's reliability and accuracy.

Feature selection is a crucial step in developing predictive models. By selecting the most relevant features from a dataset, we can avoid overfitting, reduce dimensionality, expedite the training process, and, most importantly, enhance the model's interpretability by focusing on the most influential predictors. Different feature selection methods have been proposed and widely adopted, such as stepwise regression, Lasso regression, recursive feature elimination, elastic net regression, among others. In this study, we used variance inflation factor (VIF) statistics and stepwise regression to conduct the feature selection. VIF is commonly used to assess multicollinearity in regression analysis, which quantifies how much the variance of an estimated regression coefficient increases due to the presence of multicollinearity among predictor variables. High VIF values indicate strong correlations among predictor variables. To address multicollinearity issues, we adopt an iterative approach. In each iteration, we remove the variable with the highest VIF value. This removal, in turn, tends to reduce the VIF values of the remaining variables. This process continues iteratively until all the remaining variables' VIF values fall below the recommended cut-off

threshold of 10 as suggested by previous studies (Rybarczyk and Shaker, 2021). After addressing multicollinearity, we conducted stepwise regression to further refine the model by excluding insignificant variables. Stepwise regression systematically adds or removes variables from the model based on their statistical significance. The process typically starts with an initial model, and at each step, variables are either added or removed based on specific criteria, such as Akaike information criterion (AIC), until the optimal model is achieved (Fotheringham et al., 2000).

3.4.3. Model performance evaluation

Logistic regression generally yields the probability of an event occurring. In this study, we used a threshold of 0.5 to divide the modeling results into two groups, indicating the "matched" (coded as 1) and "unmatched" (coded as 0) locations to ensure that our results are generalizable (Peng and So, 2002). Then, we compared the modelled results with the binary "ground truth" data to access modeling accuracy. We employed five-fold cross-validation to evaluate the model performance. This validation method is widely used and effective in scenarios with limited data. It randomly divides the original dataset into five subsets of equal size. The validation process is then performed five times, with each subset serving as the testing dataset while the remaining four subsets are used as the training datasets in each iteration. This approach ensures unbiased evaluation results and helps in efficiently assessing the model's performance even with a limited data sample (Li et al., 2021b). Two evaluation indices were generated in this study: accuracy and the Kappa coefficient.

4. Results

4.1. Respondent demographics

In fulfilling our first two objectives (a) and (b), we surveyed people from six towns in Central Texas from August 2021 to May 2022. A total of 286 respondents met the inclusion criteria and contributed 406 valid responses indicating perceived risk locations. The ages of respondents ranged from 19 to 81, with a mean value of 44.87. Of the respondents, 42% identified as male and 58% as female. About 43% reported a household income under $50,000. Most respondents (85%) held a valid driving license, and over half (54%) held a bachelor's or higher degree when answering the survey. Further details can be found in Appendix A.

4.2. Perceived travel risk statistics and locational relevance

The result of the Maptionnaire exercise showed that respondents were most concerned about risk near intersections, consisting of 43.3% of the sample. Road segments and POIs received a similar number of responses, making up 25.4% and 21.7% of the total. Neighborhood conditions near their chosen locales received the least number of responses, which consisted of 9.6% of the total (Fig. 4a). Participants noted that "heavy traffic volume" was the most commonly perceived risk factor, which comprised 58.9% of the sample (Fig. 4b). Secondary to this were "aggressive drivers" (38.9%). In addition, Fig. 4b indicates that the built environment also contributed to the perception of travel risk. The condition of the roadway (i.e., "poor road surface") and street lighting (i.e., "poor lighting") accounted for 16.8% and 16.3%, respectively.

Perceived risky locations as determined by the participants (red points) as well as elevated high-risk zones (pink hexagons) are displayed in Fig. 5. Among the six selected cities, Huntsville received the most responses, with 123 locations reported by our participants, making up 30.3% of the total, followed by Nolanville (87, 21.4%), Harker Heights (86, 21.2%), Copperas Cove (67, 16.5%), Caldwell (31, 7.6%), and Madisonville (12, 3.0%).

Fig. 5 clearly illustrates that perceived risk locations are not always spatially matched with observed risk locations, and likely dependent on
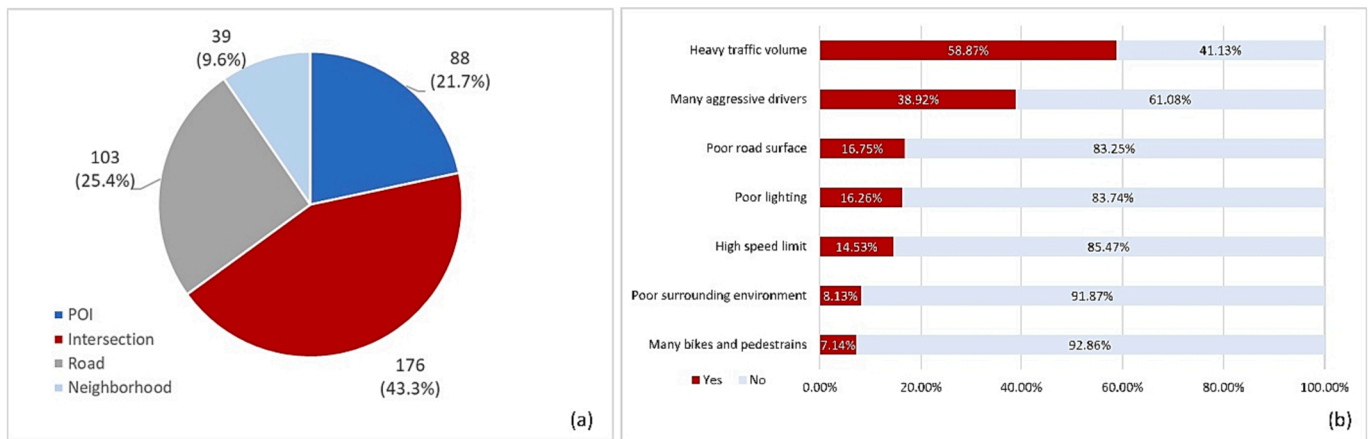
Fig. 4. Statistics of perceived risk information: (a) distribution of perceived risk types; (b) frequency of perceived risk factors.
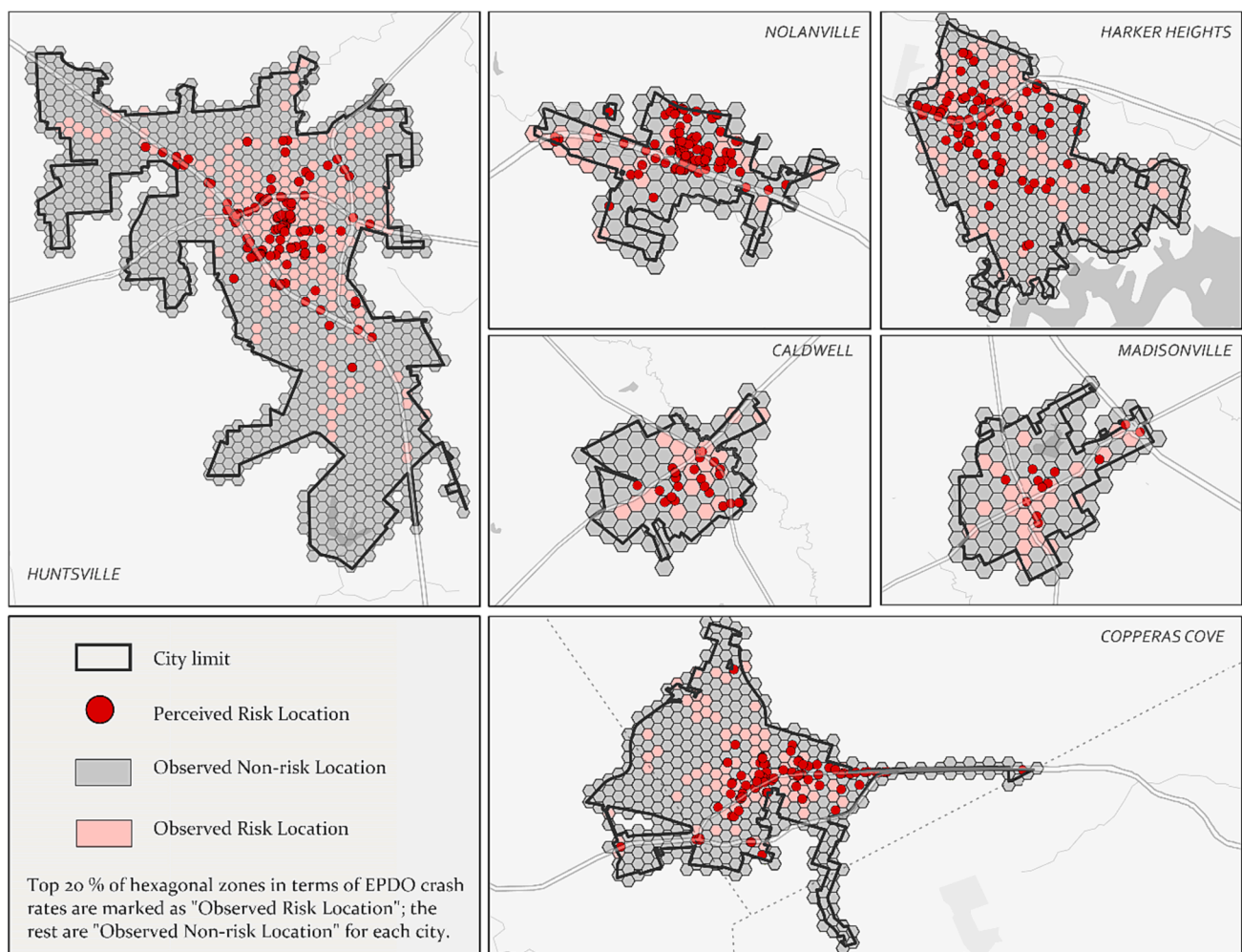

Fig. 5. Survey-based perceived risk locations and crash-based observed risk locations (grids) in six studied cities.

local conditions. Fig. 6 shows the overall and city-level matched and unmatched percentages between perceived and observed risk locations. The overall percentage of matched locations is 55.4% (Matched = 225, Unmatched = 181), which aligns with the existing findings that identifying crash sites based on perception is difficult (Duncan and Hughes, 2002). Among the six selected cities, Huntsville received the highest percentage (86.2%) of matched locations, and Nolanville received the lowest percentage of matched locations (19.5%). It implies that the

spatial relations between perceived and observed risk locations are not spatially consistent, which could be impacted by regional sociodemographic and built environmental factors.

4.3. Statistical test results

Table 3 lists the chi-square test results for a selection of categorical explanatory variables, which show a statistically significant association
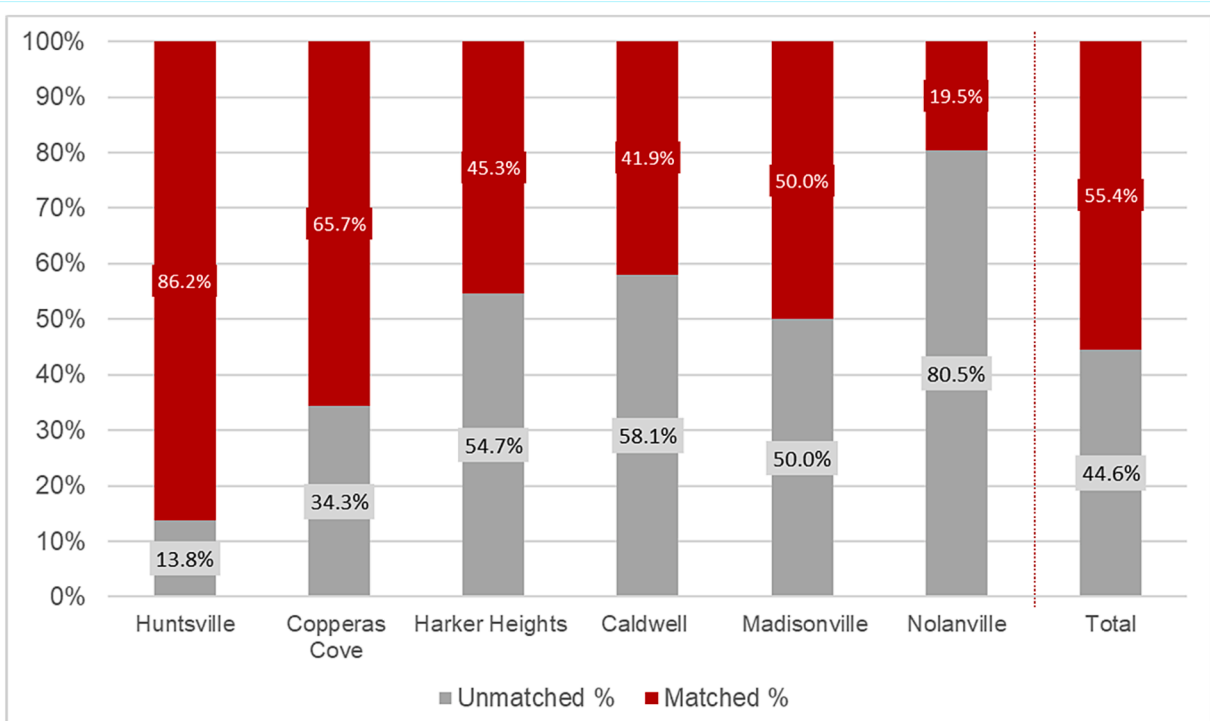
Fig. 6. The percentage of matched and unmatched perceived risk locations within each city and across all samples.

Table 3
Chi-square tests for selected categorical explanatory variables.

Variable name	Count	Unmatched	Matched
Perceived location type $\chi 2 = 41.00$; p-value = 0.000***	*POI (n = 88)*	52 (59.09%)	36 (40.91%)
	Intersection (n = 176)	48 (27.27%)	128 (72.73%)
	Road (n = 103)	54 (52.43%)	49 (47.57%)
	Neighborhood (n = 39)	27 (69.23%)	12 (30.77%)
With a valid driving license $\chi 2 = 5.11$; p-value = 0.024*	*Yes (n = 353)*	165 (46.74%)	188 (53.26%)
	No (n = 53)	16 (30.19 %)	37 (69.81%)
Total household income $\chi 2 = 14.11$; p-value = 0.003**	*Under $50 k (n = 165)*	70 (42.42%)	95 (57.58%)
	$50 k-$100 k (n = 123)	48 (39.02%)	75 (60.98%)
	Over$100 k (n = 84)	52 (61.90%)	32 (38.10%)
	Unwilling to response (n = 34)	11 (32.35%)	23 (67.65%)
City $\chi 2 = 101.24$; p-value = 0.000***	*Caldwell (n = 31)*	18 (58.06%)	13 (41.94%)
	Copperas Cove (n = 67)	23 (34.33%)	44 (65.57%)
	Harker Height (n = 86)	47 (54.65%)	39 (45.35%)
	Huntsville (n = 123)	17 (13.82%)	106 (86.18%)
	Madisonville (n = 12)	6 (50.00%)	6 (50.00%)
	Nolanville (n = 87)	70 (80.46%)	17 (19.54%)

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

with the response variable. All these variables are either perception-relevant factors or respondent's personal attributes. The results show that when a perceived location is specified as an intersection by the submitter, it is more likely to match the observed crash risk zones. The respondent's personal factors also significantly impact whether their perceived risk locations matched observed traffic risks. For example, people without a valid driving license are more sensitive to observed traffic risks, leading to a higher percentage of "matched" locations. Household income is also significantly associated with the response variable. People from high-income families (Over $100 k) had an obviously lower "matched" rate. The ratios of "matched" locations were also significantly different across cities.

Similarly, we performed the unpaired two-samples Wilcoxon tests on the continuous explanatory variables. Table 4 lists the significant variables, which include three demographic variables (*Percent of zero-car households, Percent of one-car households,* and *Percent of two-plus-car households*), three density variables (*Gross residential density, Gross population density*, and *Gross employment density*), four diversity variables (*Jobs per household, Regional diversity Employment and household entropy [trip based], Employment and household entropy*), one design variable (*Street intersection density*), two accessibility variables (*Jobs within 45 min auto travel time* and *Working age population within 45 auto travel time*), and two compound measures (*National walkability index and Smart location index*). The result indicates that the "matched" locations generally show a significantly higher value in most of these variables except *Percent of zero-car households, Percent of two-plus-car households*, and two transit accessibility variables. It implies that people's perception regarding traffic risk is more likely to align with the observed crash risks in the regions with a higher density, diversity, walkability, and location efficiency. However, the locations with a higher percentage of zero-car/two-plus-car households or higher transit accessibility usually show a lower rate of "matched" locations between perceived and observed traffic risks.

Descriptive statistics of all variables are provided in Appendix A and, a full list of statistical test results for all candidate explanatory variables is detailed in Appendix B.

4.4. Modeling results

A binary logistic regression model was implemented to distinguish

Table 4
Results of unpaired two-samples Wilcoxon tests for selected continuous explanatory variables.

Variable name	W value	p-value	Unmatched	Matched
Percent of zero-car households	23,151	0.017*	0.04 (0.06)	0.03 (0.08)
Percent of one-car households	12,671	0.000***	0.28 (0.19)	0.41 (0.26)
Percent of two-plus-car households	27,741	0.000***	0.67 (0.19)	0.55 (0.33)
Gross residential density	15,136	0.000***	0.36 (1.16)	1.62 (1.80)
Gross population density	13,214	0.000***	0.93 (3.47)	4.34 (5.16)
Gross employment density	12,787	0.000***	0.10 (1.17)	1.18 (3.87)
Jobs per household	15,389	0.000***	0.38 (0.96)	1.06 (2.15)
Employment and household entropy (trip based)	17,452	0.013*	0.59 (0.42)	0.61 (0.34)
Employment and household entropy	17,755	0.026*	0.53 (0.51)	0.59 (0.52)
Regional diversity	16,815	0.003**	0.27 (0.47)	0.46 (0.65)
Street intersection density	14,249	0.000***	17.20 (60.40)	64.80 (85.40)
Jobs within 45 min auto travel time	22,967	0.027*	1.59e4 (4.66e3)	1.51e4 (5.27e3)
Working age population within 45 auto travel time	26,607	0.000***	1.91e4 (7.41e3)	1.24e4 (8.46e3)
National walkability index	15,238	0.000***	6.67 (5.17)	7.67 (4.17)
Smart location index	14,703	0.000***	51.00 (51.10)	68.50 (32.2)

* $p < 0.05$, **$p < 0.01$, ***$p < 0.001$.
The cells under "Unmatched" and "Matched" list the median values for each variable along with the interquartile range in the parenthesis.

the "matched" and "unmatched" perceived risk locations to meet our third objective (c). Although our statistical tests identified a list of variables (listed in Section 4.3.) significantly associated with the response variable, some are collinear and cannot be included in the same model. To address this issue, we first used VIF statistics to remove intercorrelated variables. Then, we performed the stepwise feature selection to determine the optimal set of variables for the final model. Note that, to make the model replicable in other regions/cities, we excluded the variable *City* from this model even though it is statistically significant.

Table 5 displays the OR's (i.e., odds-ratios) and coefficients from the binary logistic model. We found that perceived location type—intersections ($p = 0.000$) were important with the odds (OR = 3.32) increased when perceived locations were intersections.

The model outcomes also provided evidence that roadway and neighborhood factors influenced traffic risk match rates. Of the roadway factors, street intersection density was statistically significant ($p = 0.015$). For each one unit increase in street intersection density, the odds of a successful match increased by 304.5. Although only marginally significant ($p < 0.10$, OR = 0.45), we found that road density negatively affected successful match rates. The neighborhood factor, smart location index, was statistically significant ($p = 0.003$) and exerted a strong negative influence on the successful match: for each unit increase the odds of a successful match decreased by a factor of 0.06. However, gross residential density ($p = 0.019$) and gross employment density ($p = 0.013$) exhibited a positive and significant effect. The match odds increased by a factor of 1.55 and 1.48, respectively, for each unit increase. We also found that accessibility— working age population within 45 min auto travel time ($p = 0.000$, OR = 0.07), exerted a statistically significant, albeit negative, effect on the spatial alignment between observed and perceived risk locations.

To evaluate the model's performance, we conducted a five-fold cross-validation. The detailed results can be found in Table 6, showing an average accuracy of 74.13% across all folds, with a standard deviation (SD) of 2.39%. Additionally, the average Kappa coefficient was found to be 0.48, with a SD of 0.05. These results suggest that the proposed

Table 5
Logistic regression results for distinguishing "matched" and "unmatched" perceived risk locations.

Explanatory variables	(β)	OR (95% CI)	p-value
Perception relevant factors			
Perceived location type			
Intersection (reference = POI)	1.20	**3.32 (1.74,6.37)**	**0.000*****
Road (reference = POI)	−0.08	0.92 (0.45,1.87)	0.816
Neighborhood (reference = POI)	−0.42	0.66 (0.26,1.67)	0.376
Respondent's personal attributes			
Involved in crash in past two years			
Yes (reference = No)	−0.65	0.52 (0.24,1.11)	0.091·
Retired			
Yes (reference = No)	−0.35	0.71 (0.36,1.38)	0.308
Total household income			
Under $50 k (reference = Unwilling to response)	0.07	1.08 (0.39,2.94)	0.884
$50 k-$100 k (reference = Unwilling to response)	−0.19	0.83 (0.3,2.25)	0.709
Over $100 k (reference = Unwilling to response)	−0.62	0.54 (0.19,1.5)	0.237
Number of cars within your household	0.53	1.7 (0.47,6.2)	0.419
Roadway factors			
Road density (mile/square mile)	−0.80	0.45 (0.19,1.03)	0.060 ·
Street intersection density	5.72	**304.5 (3.02,3.07e4)**	**0.015***
Built environment and location efficiency factors			
Smart location index	−2.78	**0.06 (0.01,0.4)**	**0.003****
Gross residential density	0.44	**1.55 (1.08,2.23)**	**0.019***
Gross employment density	0.39	**1.48 (1.09,2.01)**	**0.013***
Regional diversity	−0.58	0.57 (0.18,1.75)	0.324
Working age population within 45 auto travel time	−2.69	**0.07 (0.02,0.19)**	**0.000 *****

β = Estimated coefficients; OR = odds ratio, 95% confidence intervals (CI) in parentheses.
· $p < 0.1$, * $p < 0.05$, **$p < 0.01$, ***$p < 0.001$.

Table 6
Model performance assessment with five-fold cross-validation.

Evaluation Indices	Accuracy		Kappa Coefficient	
	Mean	SD	Mean	SD
Value	74.13%	±2.39%	0.48	±0.05

SD denotes standard deviation;
The value cells show the average value of five folds +/- the standard deviation.

logistic regression model performed well in distinguishing whether a perceived risk location matches the observed crash risk, achieving a satisfying level of accuracy and agreement with the actual outcomes.

5. Discussion

5.1. Key findings and policy implications

The majority of the past crash analysis research has not examined objective and perceived traffic risk together (Rankavat and Tiwari, 2020), leaving us with an unclear picture of how to accurately mitigate and forecast traffic risk. Considering the importance of perception for examining traffic risk (Rahman et al., 2021), this study has forwarded a new means to analyze risk by comprehensively integrating observed crash data with perceived traffic risks in six rural Texas rural communities. We have resultantly provided new evidence on how geography, perception, personal traits, and built environment/location efficiency affects traffic risk match rate in such areas. The methodology instituted in this research consisted of EDA, participatory GIS, spatial analysis, and binary logistic modeling. The results demonstrated that perceived traffic

risk is complicated, even for rural communities. The EDA analysis showed that the built environment and traffic volume contributed to the perception of risk. We also discovered significant geographic disparities between perceived traffic risk and high crash rate zones - signifying that risk is invariably a local phenomenon necessitating context-sensitive solutions. The magnitude of traffic risk matches was reduced in the smallest communities and suggests that progressive interventions are needed. Our binary logistic model was robust, and demonstrated that several statistically significant coefficients were important when modeling the spatial relations between perceived and observed risk locations.

Our first research objective (a) was to examine perceived traffic risk in six rural Texas communities. The majority of respondents noted that elevated risk primarily corresponded to roadway conditions. We found that intersections presented the most risky situations, and is aligned with past research (Abdel-Aty and Keller, 2005). The built environment (i.e., neighborhood and poor surrounding environment factors) had negligible impact, which is in contrast with some research pointing to the importance of neighborhood design and socio-economic status (SES) conditions as critical factors for driving risk (Morency et al., 2012; Toran Pour et al., 2017). The presence of active travelers (i.e., bicyclists and pedestrians) had little effect on perceived risk. Confirming the importance of multi-modal transportation interventions (i.e., traffic calming) as a proven means to reduce the perception of dangerous traffic situations (Alveano-Aguerrebere et al., 2018; Ewing and Dumbaugh, 2009).

In meeting our second objective (b), the EDA and spatial analysis outcomes demonstrated that traffic risk match rates varied statistically and spatially throughout our communities. Our findings showed that the match rate between perceived and observed traffic risk differed significantly within and across our study cities. The lowest match rate was in Nolanville (19.5%) and the highest was in Huntsville (86.2%). The GIS analysis showed a heterogeneous spatial relationship between perceived risk locales and high-risk zones in each community. The largest community, Huntsville, displayed numerous high-risk zones throughout the city. This is likely due to a prevalence of risky traffic environments (i.e., intersections), high population densities, and high vehicle miles traveled (VMT) due to its close proximity to Houston, all of which are conducive to higher traffic accidents. The smaller communities exhibited lower risk match rates and may in part be due to a higher portion of unreported traffic incidents. The mismatch suggests that additional traffic risk mitigation measures are needed in small underserved communities. A worthy intervention is from Park et al. (2017), where they advocated for a "community roadwatch," which utilizes digital technologies, social media, sensors, and active citizen participation to assist law enforcement officials with reducing traffic incidents. An additional prudent intervention would be a Complete Streets policy in underserved communities such as these, which would undoubtedly help reduce traffic risks and improve public health outcomes (Clifton et al., 2014).

Post Chi-square and Wilcoxon tests, we uncovered several significant factors (i.e., perceptual, roadway, and built environment/location efficiency) of explanatory variables influencing traffic risk match rates. Respondents without a valid driver's license were more sensitive to observed traffic risks, leading to more "matched" locations. Stated differently, less driving experience corresponded with more caution regarding various traffic situations and is contrast to some previous works positing the opposite (Castro et al., 2014). Similarly, respondents without recent involvement in a crash (within the past two years) were weakly associated with increased match rates ($p < 0.1$), pointing to the cautious behavior being able to better identify risky situations. Whereas those who were involved in a traffic crashes may not be as cautious due to over-confidence and not fully able to recognize dangerous traffic scenarios, hence the lowered match rate. To align perceived and real traffic risk among this group, we propose that public education campaigns be enacted in these types of communities in accordance with research from Williams (2011). By providing the public with clear data on actual risks, as well as actionable steps to take, the perceptions may

be more in line with reality. In terms of roadway, built environment, and location efficiency associations, we found several factors statistically important factors affecting match rates. This is promising, given that many past studies have identified similar factors in traffic perception (Vanderbilt, 2009; Zhang et al., 2022). Moreover, many of these factors were also discovered in our binary logistic regression outputs.

The last objective (c) in this research was to estimate the binary relationships between observed and perceived risk locations and explanatory variables representing: perceptual, personal factors, roadway conditions, and built environment/location efficiency variables. A notable perceptual factor influencing match rates was "intersection", which corresponded to high match rates, and corresponds to other studies which found similar evidence (Prajapati and Tiwari, 2013; Yang et al., 2023). We also noticed that exposure to traffic crashes weakly negated successful matches. This suggests that persons involved in a traffic crash in rural communities are generally riskier, which is in part to the transportation environment (e.g., less traffic controls, high speeds, low VMT, etc.) (Rakauskas et al., 2007) and thus is not weighed heavily when detecting dangerous travel scenarios. Relying on previous research from (Rakauskas et al., 2009), we also suggest that targeted safety education programs, versus engineering or enforcement, would be preferred by persons in rural communities such as this. The result would potentially be that people would be better able to recognize hazardous situations and elevate travel safety. Interestingly, we found that smart location index and the number of working age population within 45 min auto travel showed an inverse relationship with the response variable. This indicates that location-efficient communities, typically characterized by density, vibrancy, high walkability, and transit accessibility, are safer in terms of transportation than commonly perceived. These findings suggest that incorporating location-efficiency principles into urban planning has the potential to contribute significantly to a safer traffic environment. We also found that two neighborhood vitality factors—residential density and employment density—significantly increased traffic risk match rates. This implies that in these neighborhoods there's a heightened traffic risk awareness, which is likely due to the fact that there are more "eyes on the street" both of which have shown to moderate traffic accidents and urban stressors (Guerra et al., 2022; Jacobs, 1961). These particular neighborhoods should be modeled by planners and policy-makers and applied to other neighborhoods experiencing confusing traffic risk levels as an accident reduction strategy.

5.2. Potential safety applications based on perception data

While crash data is the primary data source in road safety assessments, it is not always available, particularly in low-income countries where their official data only capture 17% of road fatalities (World Health Organization, 2018). This study implemented a binary logistic regression model to distinguish whether perceived risk locations align with crash-based risk locations. Demonstrating an accuracy of 74.13%, tells us that it could effectively enhance the reliability of using perception data as an alternative in road safety assessments when crash data is unavailable.

Our findings also show that the "match rate" between perceived and observed traffic risks varies significantly across our studied cities. Note that this study, like most existing studies, used crash data to represent the objective crash risk. However, this approach has its limitations since most traffic incidents and near-misses are not recorded. Referring to the Heinrich pyramid, the ratio between near misses and serious injuries is around 300:1. While this ratio cannot be directly applied to safety studies, it implies that a considerable amount of traffic risk events was underrepresented or even excluded in crash-based safety studies. Recently, a growing number of studies have provided evidence that traffic incidents and near misses can be perceived by local residents and road users. Therefore, the "match rate" between perceived and observed traffic risks is important, and should be valued-added measure to

quantify the complexity of traffic risks. When the rate is low, it may imply the community has a higher portion of unreported risk events (e. g., near misses and incidents). It could alert the local authorities to reconsider whether the crash-based assessment could accurately represent crash risks, especially for small and rural towns whose crash records were less maintained.

5.3. Limitations and future work

Despite the research contributions above, we acknowledge that examining the relationship between observed and perceived risk is complicated and dependent on the methodology selected. For example, previous research has highlighted differences in perceived risk between pictures, videos, and real-world driving experiences (Charlton et al., 2014). This evidence demonstrates that the results of this study should be taken with caution.

We also acknowledge that the following limitations exist in this study. First, due to the difficulty of data collection, we only obtained 406 valid responses from six small cities in central Texas. Although the sample size is sufficient to support this pilot study, the findings from this study might be only applicable to Texas small and rural communities. As mentioned earlier, we have observed a significant variation in the "match rate" between perceived and observed traffic risks across the cities we studied. This variation suggests that different cities may have distinct crash risks (perceivable or unperceivable), which could be better captured by considering different sets of urban and road features. To achieve a higher modeling accuracy and effectively account for these variations, we suggest local authorities/researchers conduct an extensive data collection to gather more samples for their site and then build tailored models to explore the unique characteristics and crash risk factors prevalent in this location. Meanwhile, in contrast to constructing models with fixed parameters, utilizing advanced techniques such as random parameter modeling has demonstrated effectiveness in capturing potential heterogeneous effects of exploratory factors across observations (Guo et al., 2020, 2019). This can lead to improved modeling accuracy and is a promising avenue for future studies. Second, this study majorly collected data from small and rural communities. Although it could fill the knowledge gap regarding the lack of rural studies, a comparison with urban settings would be highly recommended to reveal the differences in traffic risk perception between urban and rural regions straightforwardly. Third, this study primarily explored the spatial relations between perceived risks and crash risks. But, as previously mentioned, near misses and traffic incidents account for a significant portion of traffic risk events and are directly perceivable by road users. It would be meaningful to investigate the relations between perceived risks and unreported traffic risks (e.g., near misses and incidents). Lastly, we did not account for pandemic-related effects in the crash data, which could certainly have confounded the analysis of risk factors based on pre-pandemic data from 2016 compared to data from 2020 when mobility was heavily impacted. Resultantly, the results reported here could overestimate risk perception related to traffic safety. Moving forward, it will be important for future research to re-evaluate travel behavior patterns, mode preferences, and risk perception in a post-pandemic context.

Our future efforts will focus on three perspectives: First, we will expand the study area and collect nationwide data to produce and validate more generalizable results. Second, a comparison between urban and rural settings will be added to highlight their differences and inspire rural- and urban-specific practices to leverage perception data in safety studies. Last, we will explore the association between perceived and unreported traffic risks. Currently, traffic incidents and near misses

can be captured from multiple data sources, for example, mobile apps, social media, and traffic flow data, among others. In the future study, we will leverage these multi-sourced traffic incident data to investigate how they impact people's perception of traffic risk.

6. Conclusions

Traffic risk perception directly impacts road users' traffic behaviors and can lead to different traffic risks. This study used a participatory GIS tool to collect perceived risk locations from six small towns in Central Texas and compared them to observed traffic risks. We spatially examined the relations between perceived and observed risk locations and statistically identified a set of contributing factors that could make crash-intensive areas more perceivable by road users. A binary logistic regression model was also developed, which could effectively distinguish whether perceived risk locations align with crash-based risk locations. The main findings are outlined as follows:

- First, our results highlight that road users' perceived risk locations are not always associated with high crash rates. The match rate between perceived and observed risk locations varies significantly across studied cites. A low match rate could potentially indicate more unreported traffic events (e.g., incidents and near misses) in this region.
- Second, some personal and built environment factors can significantly impact people's ability to perceive crash-intensive locations. From a personal perspective, people without a valid driving license are more likely to identify observed risk locations. Meanwhile, the perceptions of low- and middle-income road users show a higher match rate with the observed risk locations than the high-income people. From the built environment perspective, the regions with a higher density and diversity, their crash-intensive zones are more likely to align with people's perceptions. Location-efficient communities are safer in terms of transportation than commonly perceived.
- Last, through the proposed binary logistic regression model, we could effectively determine whether perceived risk locations match observed risk locations with an accuracy rate of 74.13%. This reveals the potential value of using perception data as an alternative to crash data for conducting road safety assessments.

CRediT authorship contribution statement

Xiao Li: Conceptualization, Methodology, Data curation, Formal analysis, Writing – original draft, Writing – review & editing, Visualization. **Greg Rybarczyk:** Methodology, Writing – original draft, Writing – review & editing. **Wei Li:** Conceptualization, Writing – review & editing, Funding acquisition. **Muhammad Usman:** Data curation, Writing – original draft. **Jiahe Bian:** Data curation, Writing – review & editing. **Andong Chen:** Data curation. **Xinyue Ye:** Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data that has been used is confidential.

Appendix A. Descriptive statistics of the explanatory variables.

Variable Name	Mean(SD)/Frequency(%)	Variable Name	Mean(SD)/Frequency(%)
Perceived location type		Age	
Place of interest (POI)	88 (21.67%)	Generation Z	41 (10.10%)
Intersection	176 (43.35%)	Millennials	140 (34.48%)
Road	103 (25.37%)	Generation X	84 (20.69%)
Neighborhood	39 (9.61%)	Bommers	89 (21.92%)
Perceived poor road surface		Unwilling to response	52 (12.81%)
Yes	68 (16.75%)	Marriage	
No	338 (83.25%)	Married or with a partner	283 (69.70%)
Perceived high speed limit		Divorced, Windowed or separated	52 (12.81%)
Yes	59 (14.53%)	Never Married	71 (17.49%)
No	347 (85.47%)	Total household income	
Perceived many aggressive drivers		Under $50 k	165 (40.64%)
Yes	158 (38.92%)	$50 k-$100 k	123 (30.30%)
No	248 (61.08%)	Over$100 k	84 (20.69%)
Perceived heavy traffic volume		Unwilling to response	34 (8.37%)
Yes	239 (58.87%)	Employment	
No	167 (41.13%)	Yes	237 (58.37%)
Perceived poor lighting		No	169 (41.63%)
Yes	66 (16.26%)	Retired	
No	340 (83.74%)	Yes	65 (16.01%)
Perceived poor surrounding environment		No	341 (83.99%)
Yes	33 (8.13%)	City	
No	373 (91.87%)	Caldwell	31 (7.64%)
Perceived many bikes and pedestrians		Copperas Cove	67 (16.50%)
Yes	29 (7.14%)	Harker Heights	86 (21.18%)
No	377 (92.86%)	Huntsville	123 (30.30%)
Number of cars within your household	1.48 (0.86)	Madisonville	12 (2.96%)
With a valid driving license		Nolanville	87 (21.43%)
Yes	353 (86.95%)	Road density (mile/square mile)	0.98 (0.38)
No	53 (13.05%)	Traffic exposure	2.37e7(2.96e7)
Length since licensed		Percent of zero-car households	0.05 (0.05)
<1 month	1 (0.25%)	Percent of one-car households	0.39 (0.18)
1 month to 1 year	54 (13.30%)	Percent of two-plus-car households	0.56 (0.19)
1 year to 5 years	52 (12.81%)	Gross residential density	1.38 (1.38)
Over 5 years	267 (65.76%)	Gross population density	3.49 (3.24)
Never licensed	32 (7.88%)	Gross employment density	2.37 (5.44)
Involved in crash in past two years		Jobs per household	5.47 (25.94)
Yes	49 (12.07%)	Employment and household entropy (trip based)	0.53 (0.21)
No	357 (87.93%)	Employment and household entropy	0.55 (0.25)
Received any citation in past two years		Regional diversity	0.42 (0.33)
Yes	111 (27.34%)	Street intersection density	62.24 (68.47)
No	295 (72.66%)	Jobs within 45 min auto travel time	1.46e4(4.48e3)
Gender		Working age population within 45 auto travel time	1.62e4(6.93e3)
Female	190 (46.8%)	National walkability index	7.55 (2.65)
Male	216 (53.2%)	Percent of low wage workers	0.32 (0.14)
Education		Percent of medium wage workers	0.38 (0.09)
No Bachelor's degree	165 (40.64%)	Percent of high wage workers	0.29 (0.13)
Bachelor's degree or higher	204 (50.25%)	Smart location index	57.55 (26.57)
Unwilling to response	37 (9.11%)		

Appendix B. Statistical tests for association assessment.

Chi-square test			Wilcoxon signed-rank test		
Variable name	$\chi2$	p-value	Variable name	W	p-value
Perceived location type	41.00	0.000***	Number of cars within your household	20,345	0.987
Perceived poor road surface	1.57	0.210	Road density (mile/square mile)	22,577	0.060
Perceived high speed limit	1.77	0.183	Traffic exposure	19,464	0.445
Perceived many aggressive drivers	0.01	0.929	Percent of zero-car households	23,151	0.017*
Perceived heavy traffic volume	3.75	0.053	Percent of one-car households	12,671	0.000***
Perceived poor lighting	0.94	0.333	Percent of two-plus-car households	27,741	0.000***
Perceived poor surrounding environment	2.45	0.117	Gross residential density	15,136	0.000***
Perceived many bikes and pedestrians	1.42	0.234	Gross population density	13,214	0.000***
With a valid driving license	5.11	0.024*	Gross employment density	12,787	0.000***
Length since licensed	9.334	0.053	Jobs per household	15,389	0.000***
Involved in crash in past two years	2.50	0.114	Employment and household entropy (trip based)	17,452	0.013*
Received any citation in past two years	0.31	0.578	Employment and household entropy	17,755	0.026*
Education	4.09	0.130	Regional diversity	16,815	0.003**
Gender	1.12	0.289	Street intersection density	14,249	0.000***
Age	1.54	0.820	Jobs within 45 min auto travel time	22,967	0.027*
Marriage	3.04	0.219	Working age population within 45 auto travel time	26,607	0.000***
Total household income	14.11	0.003**	National walkability index	15,238	0.000***
Employment	1.17	0.279	Percent of low wage workers	18,960	0.232
Retired	0.30	0.582	Percent of medium wage workers	22,433	0.078
City	101.24	0.000***	Percent of high wage workers	21,100	0.530
			Smart location index	14,703	0.000***

* p < 0.05, **p < 0.01, ***p < 0.001.

Chi-square tests were performed with categorical variables.
Wilcoxon signed-rank tests were performed with numerical variables.

References

Abdel-Aty, M., Keller, J., 2005. Exploring the overall and specific crash severity levels at signalized intersections. Accid. Anal. Prev. 37 (3) https://doi.org/10.1016/j.aap.2004.11.002.

AlKheder, S., Alkandriy, F., Alkhames, Z., Habeeb, M., Alenezi, R., Al Kader, A., 2022. Walkability, risk perception and safety assessment among urban college pedestrians in Kuwait. Transp. Res. Part F Traffic Psychol. Behav. 86, 10–32.

Alveano-Aguerrebere, I., Ayvar-Campos, F.J., Farvid, M., Lusk, A., 2018. Bicycle facilities that address safety, crime, and economic development: Perceptions from Morelia, Mexico. Int. J. Environ. Res. Public Health 15, 1. https://doi.org/10.3390/ijerph15010001.

American Association of State Highway and Transportation Officials, 2010. Highway Safety Manual, 1st Edition, 1st ed.

Austin Capital Advisors, 2021. Texas Triangle: An Emerging Mega Region [WWW Document]. accessed 3.2.23. https://www.austincapitaladvisors.com/texas-triangle.

Blatt, J., Furman, S.M., 1998. Residence location of drivers involved in fatal crashes. Accid. Anal. Prev. 30 (6), 705–711. https://doi.org/10.1016/S0001-4575(98)00014-1.

Cabrera-Arnau, C., Prieto Curiel, R., Bishop, S.R., 2020. Uncovering the behaviour of road accidents in urban areas. Uncovering the behaviour of road accidents in urban areas. 7 (4), 191739.

Cai, Q., Abdel-Aty, M., Lee, J., Eluru, N., 2017. Comparative analysis of zonal systems for macro-level crash modeling. J. Saf. Res. 61, 157–166.

Castro, C., Padilla, J.L., Roca, J., Benítez, I., García-Fernández, P., Estévez, B., López-Ramón, M.F., Crundall, D., 2014. Development and Validation of the Spanish Hazard Perception Test. Traffic Inj. Prev. 15 (8), 817–826.

Charlton, S.G., Starkey, N.J., Perrone, J.A., Isler, R.B., 2014. What's the risk? A comparison of actual and perceived driving risk. Transport. Res. F: Traffic Psychol. Behav. 25, 50–64.

Chimba, D., Musinguzi, A., Kidando, E., 2018. Associating pedestrian crashes with demographic and socioeconomic factors. Case Studies in Transport Policy 6 (1), 11–16. https://doi.org/10.1016/j.cstp.2018.01.006.

Chmielewski, J., Kempa, J., 2020. Hexagonal Zones in Transport Demand Models. KnE Engineering. 10.18502/keg.v5i6.7025.

Clifton, K., Bronstein, S., Morrissey, S., 2014. The Path to Complete Streets in Underserved Communities: Lessons from U.S. Case Studies [WWW Document]. URL https://activelivingresearch.org/sites/activelivingresearch.org/files/Clifton_Path_to_Complete_Streets_Underserved_Communities_Oct2014.pdf (accessed 3.2.23).

Deery, H.A., 1999. Hazard and Risk Perception among Young Novice Drivers. J. Saf. Res. 30 (4), 225–236.

DeJoy, D.M., 1989. The optimism bias and traffic accident risk perception. Accid. Anal. Prev. 21 (4), 333–340.

Duncan, C., Hughes, R.G., 2002. The perceived effectiveness of sidewalks and other pedestrian safety treatments as countermeasures for hit-along-roadway crashes [WWW Document]. URL https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.194.9728&rep=rep1&type=pdf (accessed 9.16.22).

Ewing, R., Dumbaugh, E., 2009. The built environment and traffic safety: A review of empirical evidence. J. Plan. Lit. 23 (4), 347–367.

Ferenchak, N.N., 2022. U.S. Vision Zero Cities: modal fatality trends and strategy effectiveness. Transportation Letters 1–12. 10.1080/19427867.2022.2116673.

Fotheringham, A.S., Brunsdon, C., Charlton, M., 2000. Quantitative geography: Perspectives on spatial data analysis. Sage Publications Ltd.

Ghadiri, M., Rassafi, A.A., Mirbaha, B., 2019. The effects of traffic zoning with regular geometric shapes on the precision of trip production models. J. Transp. Geogr. 78, 150–159.

Guerra, E., Dong, X., Kondo, M., 2022. Do Denser Neighborhoods Have Safer Streets? Population Density and Traffic Safety in the Philadelphia Region. J. Plan. Educ. Res. 42, 4. https://doi.org/10.1177/0739456X19845043.

Guo, Y., Li, Z., Liu, P., Wu, Y., 2019. Modeling correlation and heterogeneity in crash rates by collision types using full bayesian random parameters multivariate Tobit model. Accid. Anal. Prev. 128, 164–174. https://doi.org/10.1016/j.aap.2019.04.013.

Guo, Y., Sayed, T., Essa, M., 2020. Real-time conflict-based Bayesian Tobit models for safety evaluation of signalized intersections. Accid. Anal. Prev. 144, 105660 https://doi.org/10.1016/j.aap.2020.105660.

Huang, Y., Wang, X., Patton, D., 2018. Examining spatial relationships between crashes and the built environment: A geographically weighted regression approach. J. Transp. Geogr. 69, 221–233.

International Transport Forum, 2018. Road Safety Annual Report 2018 [WWW Document]. accessed 3.1.23. https://www.itf-oecd.org/sites/default/files/docs/irtad-road-safety-annual-report-2018_2.pdf.

Jacobs, J., 1961. The Death and Life of Great American Cities. Book Unpublished resources, Randoms House, New York.

Karim, M.R., 1992. Traffic Accidents in a University Environment. ITE Journal 62, 7.

Kononov, J., Allery, B., Znamenacek, Z., 2007. Safety planning study of urban freeways: Proposed methodology and review of case history. Transp. Res. Rec. 2019 (1), 146–155.

LaPlante, J., McCann, B., 2008. Complete Streets: We Can Get There from Here. Ite Journal-institute of Transportation Engineers 78, 24–28.

Lee, S.M., Conway, T.L., Frank, L.D., Saelens, B.E., Cain, K.L., Sallis, J.F., 2017. The Relation of Perceived and Objective Environment Attributes to Neighborhood Satisfaction. Environ. Behav. 49 (2), 136–160.

Lee, G., Park, Y., Kim, J., Cho, G.-H., 2016. Association between intersection characteristics and perceived crash risk among school-aged children. Accid. Anal. Prev. 97, 111–121.

Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M., Cai, Q., 2018. Analysis of crash proportion by vehicle type at traffic analysis zone level: A mixed fractional split multinomial logit modeling approach with spatial effects. Accid. Anal. Prev. 111, 12–22.

Leviäkangas, P., 1998. Accident risk of foreign drivers - The case of Russian drivers in south-eastern Finland. Accid. Anal. Prev. 30 (2), 245–254.

Li, X., Mousavi, S.M., Dadashova, B., Lord, D., Wolshon, B., 2021a. Toward a crowdsourcing solution to identify high-risk highway segments through mining driving jerks. Accid. Anal. Prev. 155, 106101 https://doi.org/10.1016/j.aap.2021.106101.

Li, X., Ning, H., Huang, X., Kang, Y., Dadashova, B., Ma, A., 2021b. Urban infrastructure audit: an effective protocol to digitize signalized intersections by mining street view images. Cartogr. Geogr. Inf. Sci. 49 (1), 32–49. https://doi.org/10.1080/15230406.2021.1992299.

Li, X., Yu, S., Huang, X., Dadashova, B., Cui, W., Zhang, Z., 2022. Do underserved and socially vulnerable communities observe more crashes? A spatial examination of social vulnerability and crash risks in Texas. Accid. Anal. Prev. 173, 106721 https://doi.org/10.1016/J.AAP.2022.106721.

Machado-León, J.L., de Oña, J., de Oña, R., Eboli, L., Mazzulla, G., 2016. Socio-economic and driving experience factors affecting drivers' perceptions of traffic crash risk. Transp Res Part F Traffic Psychol Behav 37, 41–51.

Morency, P., Gauvin, L., Plante, C., Fournier, M., Morency, C., 2012. Neighborhood social inequalities in road traffic injuries: The influence of traffic volume and road design. Am. J. Public Health 102 (6), 1112–1119.

Nie, K., Wang, Z., Du, Q., Ren, F., Tian, Q., 2015. A network-constrained integrated method for detecting spatial cluster and risk location of traffic crash: A case study from Wuhan, China. Sustainability (Switzerland) 7 (3), 2662–2677. https://doi.org/10.3390/su7032662.

Noland, R.B., 1995. Perceived risk and modal choice: Risk compensation in transportation systems. Accid. Anal. Prev. 27 (4), 503–521.

Nordfjærn, T., Rundmo, T., 2009. Perceptions of traffic risk in an industrialised and a developing country. Transp Res Part F Traffic Psychol Behav 12 (1), 91–98.

Park, S., Ilincai, E.S., Oh, J., Kwon, S., Mizouni, R., Lee, U., 2017. Facilitating pervasive community policing on the road with mobile roadwatch. In: in: Conference on Human Factors in Computing Systems - Proceedings. https://doi.org/10.1145/3025453.3025867.

Peng, C.-Y., So, T.-S., 2002. Logistic Regression Analysis and Reporting: A Primer. Underst. Stat. 1 (1), 31–70.

Prajapati, P., Tiwari, G., 2013. Study of Relation between Actual and Perceived Crash Risk. Procedia Soc. Behav. Sci. 104, 1095–1104.

Rahman, M.R., Poudel, N., Singleton, P.A., 2021. Multimodal traffic safety concerns in a university population. Transp Res Part F Traffic Psychol Behav 80, 424–435.

Rakauskas, M., Ward, N., Gerberich, S., Alexander, B., 2007. Rural and Urban Safety Cultures: Human-Centered Interventions Toward Zero Deaths in Rural Minnesota.

Rakauskas, M.E., Ward, N.J., Gerberich, S.G., 2009. Identification of differences between rural and urban safety cultures. Accid. Anal. Prev. 41 (5), 931–937. https://doi.org/10.1016/j.aap.2009.05.008.

Ramsey, K., Bell, A., 2014. Smart Location Database: Version 2.0 User Guide [WWW Document]. URL https://www.epa.gov/sites/default/files/2014-03/documents/sld_userguide.pdf (accessed 3.2.23).

Rankavat, S., Tiwari, G., 2016. Pedestrians risk perception of traffic crash and built environment features – Delhi, India. Saf. Sci. 87, 1–7. https://doi.org/10.1016/J.SSCI.2016.03.009.

Rankavat, S., Tiwari, G., 2020. Influence of actual and perceived risks in selecting crossing facilities by pedestrians. Travel Behav. Soc. 21, 1–9.

Rhee, K.-A., Kim, J.-K., Lee, Y.-I., Ulfarsson, G.F., 2016. Spatial regression analysis of traffic crashes in Seoul. Accid. Anal. Prev. 91, 190–199. https://doi.org/10.1016/j.aap.2016.02.023.

Rybarczyk, G., Shaker, R.R., 2021. Predicting bicycle-on-board transit choice in a university environment. Sustainability (Switzerland) 13, 2. https://doi.org/10.3390/su13020512.

Rybarczyk, G., Banerjee, S., Starking-Szymanski, M.D., Shaker, R.R., 2018. Travel and us: the impact of mode share on sentiment using geo-social media and GIS. Journal of Location Based Services 12 (1), 40–62.

Schneider, R.J., Ryznar, R.M., Khattak, A.J., 2004. An accident waiting to happen: a spatial approach to proactive pedestrian planning. Accid. Anal. Prev. 36 (2), 193–211. https://doi.org/10.1016/S0001-4575(02)00149-5.

Song, J.J., Ghosh, M., Miaou, S., Mallick, B., 2006. Bayesian multivariate spatial models for roadway traffic crash mapping. J. Multivar. Anal. 97 (1), 246–273.

Subasi, A., 2020. Practical Machine Learning for Data Analysis Using Python. Practical Machine Learning for Data Analysis Using Python. https://doi.org/10.1016/B978-0-12-821379-7.00008-4.

Svenson, J.E., Spurlock, C., Nypaver, M., 1996. Factors Associated With the Higher Traumatic Death Rate Among Rural Children. Ann. Emerg. Med. 27 (5), 625–632. https://doi.org/10.1016/S0196-0644(96)70167-1.

Thakali, L., Kwon, T.J., Fu, L., 2015. Identification of crash hotspots using kernel density estimation and kriging methods: a comparison. Journal of Modern Transportation 23 (2), 93–106.

The U.S. National Highway Safety Administration, 2019. Crash Rate Calculations [WWW Document]. URL https://safety.fhwa.dot.gov/local_rural/training/fhwasa1109/app_c.cfm (accessed 7.5.20).

The United States Office of Management and Budget, 2022. Delineating Metropolitan and Micropolitan Statistical Areas [WWW Document]. accessed 3.2.23. https://www.census.gov/programs-surveys/metro-micro/about.html.

Toran Pour, A., Moridpour, S., Tay, R., Rajabifard, A., 2017. Neighborhood Influences on Vehicle-Pedestrian Crash Severity. J. Urban Health 94 (6), 855–868.

United Nations, 2019. Road Safety Strategy [WWW Document]. accessed 9.6.22. https://www.un.org/sites/un2.un.org/files/2020/09/road_safety_strategy_booklet.pdf.

United States Census Bureau, 2021. QuickFacts - Texas [WWW Document]. accessed 9.15.22. https://www.census.gov/quickfacts/TX.

Vanderbilt, T., 2009. Traffic: Why We Drive the Way We Do (and What It Says About Us). 10.17077/drivingassessment.1294.

von Stülpnagel, R., Lucas, J., 2020. Crash risk and subjective risk perception during urban cycling: Evidence for congruent and incongruent sources. Accid. Anal. Prev. 142, 105584.

Wang, B., Hensher, D.A., Ton, T., 2002. Safety in the road environment: A driver behavioural response perspective. Transportation (Amst) 29, 3. https://doi.org/10.1023/A:1015661008598.

Williams, A.F., 2011. Teenagers' Licensing Decisions and Their Views of Licensing Policies: A National Survey. Traffic Inj. Prev. 12 (4), 312–319. https://doi.org/10.1080/15389588.2011.572100.

World Health Organization, 2018. Global status report on road safety 2018 [WWW Document]. accessed 3.1.23. https://www.who.int/publications/i/item/9789241565684.

World Health Organization, 2020. The top 10 causes of death [WWW Document]. accessed 3.1.23. https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death.

Yang, L., Luo, X., Zuo, Z., Zhou, S., Huang, T., Luo, S., 2023. A novel approach for fine-grained traffic risk characterization and evaluation of urban road intersections. Accid. Anal. Prev. 181, 106934 https://doi.org/10.1016/J.AAP.2022.106934.

Yang, H., Ma, Q., Wang, Z., Cai, Q., Xie, K., Yang, D.i., 2020. Safety of micro-mobility: Analysis of E-Scooter crashes by mining news reports. Accid. Anal. Prev. 143, 105608.

Yao, S., Loo, B.P.Y., Yang, B.Z., 2016. Traffic collisions in space: four decades of advancement in applied GIS. Ann. GIS 22 (1), 1–14.

Yu, H., Liu, P., Chen, J., Wang, H., 2014. Comparative analysis of the spatial analysis methods for hotspot identification. Accid. Anal. Prev. 66, 80–88.

Zhang, D., Chen, F., Zhu, J., Wang, C., Cheng, J., Zhang, Y., Bo, W.u., Zhang, P., 2022. Research on drivers' hazard perception in plateau environment based on visual characteristics. Accid. Anal. Prev. 166, 106540.