SJSU SAN JOSÉ STATE UNIVERSITY

MINETA TRANSPORTATION INSTITUTE

# Equitable Estimation of Accurate High Injury Networks (HINs) for Vulnerable Road Users

Anurag Pande, PhD
Anandamayee Majumdar, PhD
Lucas Roman Kantorowski


California State University Transportation Consortium

CAL POLY SAN LUIS OBISPO

# MINETA TRANSPORTATION INSTITUTE

Founded in 1991, the Mineta Transportation Institute (MTI), an organized research and training unit in partnership with the Lucas College and Graduate School of Business at San José State University (SJSU), increases mobility for all by improving the safety, efficiency, accessibility, and convenience of our nation's transportation system. Through research, education, workforce development, and technology transfer, we help create a connected world. MTI leads the California State University Transportation Consortium (CSUTC) funded by the State of California through Senate Bill 1 and the Climate Change and Extreme Events Training and Research (CCEETR) Program funded by the Federal Railroad Administration. MTI focuses on three primary responsibilities:

### Research

MTI conducts multi-disciplinary research focused on surface transportation that contributes to effective decision making. Research areas include: active transportation; planning and policy; security and counterterrorism; sustainable transportation and land use; transit and passenger rail; transportation engineering; transportation finance; transportation technology; and workforce and labor. MTI research publications undergo expert peer review to ensure the quality of the research.

### Education and Workforce Development

To ensure the efficient movement of people and goods, we must prepare the next generation of skilled transportation professionals who can lead a thriving, forward-thinking transportation industry for a more connected world. To help achieve this, MTI sponsors a suite of workforce development and education opportunities. The Institute supports educational programs offered by the Lucas Graduate School of Business: a Master of Science in Transportation Management, plus graduate certificates that include High-Speed and Intercity Rail Management and Transportation Security Management. These flexible programs offer live online classes so that working transportation professionals can pursue an advanced degree regardless of their location.

### Information and Technology Transfer

MTI utilizes a diverse array of dissemination methods and media to ensure research results reach those responsible for managing change. These methods include publication, seminars, workshops, websites, social media, webinars, and other technology transfer mechanisms. Additionally, MTI promotes the availability of completed research to professional organizations and works to integrate the research findings into the graduate education program. MTI's extensive collection of transportation-related publications is integrated into San José State University's world-class Martin Luther King, Jr. Library.

Report 26-08

# Equitable Estimation of Accurate High-Injury Networks (HINs) for Vulnerable Road Users

Anurag Pande, PhD

Anandamayee Majumdar, PhD

Lucas Roman Kantorowski

April 2026

A publication of the
Mineta Transportation Institute
Created by Congress in 1991

College of Business
San José State University
San José, CA 95192-0219

# T TECHNICAL REPORT DOCUMENTATION PAGE

| | | |
|---|---|---|
| **1. Report No.**<br>26-08 | **2. Government Accession No.** | **3. Recipient's Catalog No.** |

| | |
|---|---|
| **4. Title and Subtitle**<br>Equitable Estimation of Accurate High Injury Networks (HINs) for Vulnerable Road Users | **5. Report Date**<br>April 2026<br><br>**6. Performing Organization Code** |
| **7. Author**<br>Anurag Pande, PhD 0000-0002-3456-7932<br>Anandamayee Majumdar, PhD 0000-0001-7793-4037<br>Lucas Roman Kantorowski 0009-0009-3284-1166 | **8. Performing Organization Report**<br>CA-MTI-2459 |
| **9. Performing Organization Name and Address**<br>Mineta Transportation Institute<br>College of Business<br>San José State University<br>San José, CA 95192-0219 | **10. Work Unit No.**<br><br>**11. Contract or Grant No.**<br>SB1-SJAUX_2023-26 |
| **12. Sponsoring Agency Name and Address**<br>State of California SB1 2017/2018<br>Trustees of the California State University<br>Sponsored Programs Administration<br>401 Golden Shore, 5th Floor<br>Long Beach, CA 90802 | **13. Type of Report and Period Covered**<br><br>**14. Sponsoring Agency Code** |

**15. Supplemental Notes**
10.31979/mti.2025.2459

**16. Abstract**
To improve traffic safety, communities first need to know where serious crashes are actually happening. High-Injury Networks are designed to identify these locations, but they are usually built using only police-reported crash data. This research asks: Are we missing crashes—and injuries—by relying on police data alone? The research team first demonstrates that data from Emergency Medical Services and those from official databases differ substantially, and neither data set captures the full extent of collisions in a community. Following that analysis, the research team concluded that a new data source may help complete this. Toward that end, the research team scraped data from PulsePoint in San Francisco to identify potential traffic collisions reported to 911 operators. The 911 call data for San Francisco, when compared with official city crash data sources, shows that several traffic incidents reported on 911 calls do not appear in the city's official database. Statistical analysis of data from both sources vs. those found in only one of the two reveals patterns by location. Locations in police districts with lower population density (and larger geographical areas) had more 911 call-reported incidents that did not appear in the official database. The demographics of census tracts of the incident's reported location, such as income, race, and education levels, did not appear to be statistically significant. Based on the findings, the research team provides a framework for complementing collision data with alternative sources beyond the police records in future Vision Zero efforts. The research project also resulted in a process that allows the team to continuously add to the scraped 911 call data, enabling this analysis to continue beyond what is presented in this report. When serious injuries are invisible in the data, they are invisible in safety planning. Integrating and using all available data is critical to ensuring that Vision Zero strategies reflect real-world injury risk and deliver meaningful, life-saving outcomes.

| | |
|---|---|
| **17. Key Words**<br>Transportation safety, crash data, vulnerable road users, demographics, spatial analysis. | **18. Distribution Statement**<br>No restrictions. This document is available to the public through The National Technical Information Service, Springfield, VA 22161. |

| | | | |
|---|---|---|---|
| **19. Security Classif. (of this report)**<br>Unclassified | **20. Security Classif. (of this page)**<br>Unclassified | **21. No. of Pages**<br>48 | **22. Price** |

Form DOT F 1700.7 (8-72)

DOI: 10.31979/mti.2026.2459

Mineta Transportation Institute
College of Business
San José State University
San José, CA 95192-0219

Tel: (408) 924-7560
Fax: (408) 924-7565
Email: mineta-institute@sjsu.edu

transweb.sjsu.edu/research/2460

# Acknowledgments

The authors want to thank Charlotte Jackson and her team at Data Driven Streets for their support in scraping 911 call data from PulsePoint. We also want to thank the San Luis Obispo Council of Governments and Kittleson and Associates for providing EMS (Emergency Medical Services) data for the city of San Luis Obispo. Mr. Ryan M. Rosander from the County of San Luis Obispo Emergency Medical Services Agency (EMSA), the local agency responsible for administering emergency medical services in the county, provided valuable context for the data and patterns observed. Mr. John Clamon, a graduate student at Cal Poly, helped with revisions to the draft report. The authors thank Lisa Rose and Editing Press for editorial services, as well as MTI staff Project Assistant, Rajeshwari Rajesh, and Graphic Design Assistant, Katerina Earnest.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# List of Abbreviations

| | |
|---|---|
| HIN | High-injury Network |
| EMS | Emergency Medical Services |
| VZ | Vision Zero |
| VRU | Vulnerable Road Users |
| SWITRS | Statewide Integrated Traffic Records System |
| TIMS | Transportation Injury Mapping System |

# Executive Summary

This research investigates the completeness of traditional police-reported crash data used for the development of high-injury networks (HINs). It also examines whether alternative data sources can provide a more accurate and representative picture of traffic crash risk. Specifically, the study examines two sources of data, Emergency Medical Services (EMS) data and 911 call data, for their ability to supplement police crash records to improve the estimation of HINs for implementation of the Vision Zero (VZ) policy framework.

Vision Zero (VZ) is a policy framework grounded in a moral and public health principle that no loss of life in traffic is acceptable. The framework seeks to eliminate traffic fatalities and serious injuries through a systemic approach to roadway safety informed by data. High-injury networks (HINs), the subset of roadways where the greatest concentration of severe and fatal crashes occurs, are a key element of VZ planning. HINs are traditionally derived from police-reported collision data. However, growing evidence suggests that these data may be incomplete, particularly for crashes involving vulnerable road users (VRUs), including pedestrians and bicyclists. Incomplete or biased crash data can lead to misidentification of high-risk locations and misallocation of safety improvement dollars.

The research was conducted in two phases. First, EMS collision data from the City of San Luis Obispo were compared with police-reported crash data from the Statewide Integrated Traffic Records System (SWITRS). This comparison revealed systematic discrepancies between the two datasets. Numerous bicycle- and pedestrian-involved injury crashes recorded by EMS, particularly near the Cal Poly campus and surrounding streets, did not appear in police crash records. Overall, the findings demonstrated that neither dataset alone provides a complete representation of traffic injuries and that reliance on police-reported data may systematically undercount certain types of VRU incidents.

In the second phase, the study analyzed 911 call data scraped in real time from PulsePoint. PulsePoint is an app that notifies CPR-trained persons of cardiac arrest events in their immediate vicinity for communities working with PulsePoint. For participating communities, they also provide data for live emergency calls with associated details such as the type of emergency. We used this PulsePoint data for the City of San Francisco and compared these records with the City's official police-reported traffic collision database. The analysis showed that a substantial share of traffic collisions reported via 911 calls were not part of the official crash database. Furthermore, the degree of overlap between 911 call data and police-reported crashes varied significantly across geographical areas represented by police districts. Denser, more institutional districts such as the Mission, Central, Northern, and the Tenderloin exhibited higher matching rates, while more residential districts, including Richmond, Ingleside, Taraval, Southern, and Bayview, showed considerably lower matching rates.

Factors such as population density, time of occurrence, and incident type influence whether a 911-reported collision is captured in the official crash database. Neighborhood-level demographic characteristics, including income and racial composition, were not statistically significant in the estimated models. These results suggest that spatial context (residential vs. commercial-heavy areas; population density) and reporting dynamics (e.g., due to time of incident occurrence) play a larger role than socioeconomic characteristics (e.g., ethnic makeup of the population in the neighborhood) alone in determining data completeness.

Overall, the findings substantiate that police-reported crash data alone do not fully capture the distribution of traffic injuries, particularly for vulnerable road users. Integrating EMS and 911 call data can potentially provide a more comprehensive and timely understanding of the risk of severe injuries and fatalities. A framework for the next steps agencies can take based on the findings of this work is provided in the report. For agencies implementing VZ policies, using these complementary data sources can lead to more accurate identification of HINs, better prioritization of safety investments, and improved equity outcomes.

# 1. Introduction

Vision Zero (VZ), a policy framework to reduce roadway traffic fatalities to zero, can provide strategic direction for systemic efforts to achieve safer, more equitable mobility for all road users. Most U.S. cities have not achieved the targeted decline in fatalities and injuries even after adopting VZ as the guiding policy framework. A critical task for communities considering VZ implementation is to conduct a systemic analysis to identify locations with the highest risk of dangerous traffic collisions and similar roadway characteristics that lend themselves to improvement. These sets of locations are known as high-injury networks (HINs). This research effort addresses critical questions from the literature relevant to accurately identifying HINs and, hence, to the VZ policy goals.

## 1.1 Background

Vision Zero (VZ) policies aim to alleviate the public health crisis of traffic crashes. First implemented in Sweden in the 1990s, Vision Zero as a policy framework has proved successful across Europe, and it is gaining momentum in large and even small-to-medium-sized American cities (including SLO). However, barring a few exceptions, such as Boston, MA ("Bucking National Trends, City of Boston Marks Progress on 'Vision Zero'" 2021), and Fremont, CA ("Fremont Vision Zero 2020 | City of Fremont Official Website" n.d.), VZ adoption has not resulted in dramatic declines in serious traffic crashes; in fact, collisions involving VRUs have been rising in the U.S.

Identifying high-injury networks (HINs) and applying countermeasures are key to the successful implementation of VZ, and its effectiveness depends on accurate historical crash data. Traditionally, HIN estimation and validation processes are based on police crash and injury reports. In terms of reliable high-injury surface street networks, there is a significant concern that the traditional police-reports-based crash data are not fully capturing the extent of serious incidents and injuries. Several VRU advocates point to the anecdotal evidence (*Washington Post* n.d.), but there is also published literature documenting the issues (Noland et al. 2017). Noland et al. (2017) highlighted the problems in the classification, reporting, and analysis of pedestrian deaths in New Jersey. In a 1999 study, Elvik and Mysen showed the collision reporting to be incomplete at all levels of severity. They observed that the more severe the injury, the more likely the event was to be reported. Their results also showed that only 49% of hospital-treated injury collisions in the U.S. are registered in official road collision statistics. Saxton (Saxton, 2018) noted that reporting tends are highest for car occupants and lowest for one of the VRU groups, i.e., cyclists (Elvik and Mysen 1999).

This study aims to address the critical challenge of developing an equitable and accurate method for developing high-injury networks (HINs) to implement Vision Zero policies. The challenge is especially acute for injuries[1] (VRUs). Portland's High Crash Network methodology notes that

[1] VRUs typically include bicyclists and pedestrians but may include others, e.g., people using wheelchairs.

using police reports often results in a significant delay (sometimes more than a year) (“High Crash Network Streets and Intersections", Portland.Gov n.d.), even when data are available in a timely manner. An analysis by Pulugurtha (Pulugurtha) also noted the need for a more refined and less subjective process to identify HINs. They specifically noted issues with crash-rate-based measures used to identify HINs. This method often produces somewhat counterintuitive results in that pedestrian amenities such as marked crosswalks are frequently (Quistberg).

Therefore, data that incorporate traffic collisions as medical emergencies and include crashes where people left the scene, and collisions that were not recorded by the police department (Calma and Jackson 2021), are needed for agencies to have a complete picture of the crashes. Without such a dataset on where VRUs are involved in collisions, it is impossible to make truly informed decisions to achieve VZ policy goals. In this study, we used EMS data from San Luis Obispo County, California, and 911 call data from San Francisco to demonstrate to what extent police-reported data may reflect the collision experience on city streets.

## 1.2 Research Questions

The reliability of the underlying data used for the identification of HINs needs to be investigated to address the following questions:

- Are VRU-involved crashes under-reported and/or reported with a systemic bias (e.g., affected by the affluence of the crash location neighborhood)?

- Are such systemic biases preventing the decision-makers from identifying threats to safety and effectively directing resources toward locations and facilities?

## 1.3 Report Organization

The next section presents a detailed review of the literature that discusses the biases in the collision data and potential ways to address them. Based on the literature, Chapter 3 presents the results of the analysis of EMS data, demonstrating that in urban networks, neither police-reported data nor EMS data provide a full picture of all injury collisions. Chapter 4 provides a methodology for capturing 911 call data to provide a more complete picture of all collisions, using San Francisco as a case study. The paper also analyzes differences between incidents reported in both police-reported and 911 call data and those found only in the PulsePoint data. The final chapter presents the conclusions and future scope of this work.

# 2. Literature Review

Vision Zero (VZ) policies were developed to address what has long been recognized as a global public health crisis of traffic crashes resulting in severe injuries and fatalities. Originating in Sweden in the 1990s (*Vision Zero – a History* 2019), Vision Zero has since been adopted by governments around the world, including several U.S. jurisdictions. The philosophy is straightforward: no loss of life on the road is acceptable, and system design should anticipate human error and mitigate its consequences. Despite the widespread appeal of Vision Zero as a policy framework, implementation outcomes in the United States range from poor to mixed. Many cities, from large metropolitan areas such as Boston, Massachusetts, to smaller jurisdictions such as Fremont, California, have adopted Vision Zero frameworks and demonstrated measurable progress in some safety metrics ("Bucking National Trends, City of Boston Marks Progress on 'Vision Zero'" 2021; "Fremont Vision Zero," City of Fremont Official Website n.d.). However, the overall rate of serious crashes, especially those involving pedestrians, has continued to rise nationally ("Pedestrian Traffic Fatalities by State" n.d.).

## 2.1 High-Injury Networks: The Data Challenge

A major challenge in realizing VZ goals is the reliability of the data that guides safety planning. Identifying high-injury networks (HINs), the subsets of the road network where the largest fraction of crashes with serious injuries and fatalities occur, is central to Vision Zero's data-driven approach. The accuracy of these networks depends on the completeness of the underlying crash data. Police-reported crash data form the basis of most traffic safety monitoring and policy evaluation in the United States. This reliance on police-reported crash data has significant limitations, particularly for incidents involving vulnerable road users (VRUs), including pedestrians, cyclists, and micromobility users.

Police crash reports are designed for enforcement and legal documentation, not for public health surveillance. Unsurprisingly, they systematically omit many crash victims, particularly those with non-fatal injuries, vulnerable road users, and incidents occurring on private property or outside the scope of routine police response. A growing body of evidence shows that these data systems capture only a fraction of the true injury burden (Noland et al. 2017; Elvik and Mysen 1999). Noland et al. (2017) documented widespread misclassification and incomplete reporting of pedestrian deaths in New Jersey, while Elvik and Mysen (1999) found that only 49% of hospital-treated crash injuries were included in official U.S. road safety statistics. The underreporting was particularly acute for non-fatal incidents involving automobiles as well as those involving VRUs. Reporting completeness tends to be highest among vehicle occupants and lowest among pedestrians and cyclists, precisely the groups most vulnerable on the nation's roadways.

In addition to underreporting, police crash data are often delayed, inconsistent, and subject to local variations in reporting practices. For example, Portland's High Crash Network methodology notes that complete crash data can take more than a year to finalize ("High Crash Network Streets and

Intersections," Portland.Gov n.d.). Even after data are made available, extensive cleaning, standardization, and post-processing are required before meaningful analysis can occur (Saxton 2018). These challenges hinder timely and accurate identification of high-risk corridors and complicate efforts to monitor progress toward Vision Zero objectives.

Other studies (Pulugurtha et al. 2007; Quistberg et al. 2015) underscored that the use of crash-rate-based measures to identify high-injury networks can yield counterintuitive and counterproductive results. For instance, pedestrian amenities such as marked crosswalks can appear statistically correlated with higher crash rates, reflecting not greater risk but rather higher pedestrian volumes or more accurate reporting. These findings highlight the need for more comprehensive and context-sensitive data sources.

## 2.2 High-Injury Networks: Integrating Data Sources

Recognizing these limitations, recent research has focused on integrating police reports with hospital, trauma registry, and emergency medical services (EMS) data to improve completeness and reliability. These multi-source linkage approaches have revealed that official police data substantially underestimate the true burden of traffic injuries.

Soltani et al. (2022) linked police crash reports with hospital discharge data, EMS run logs, and medical examiner records to create a unified Vision Zero injury surveillance database. The resulting dataset captured thousands of additional crash injuries that never appeared in police files. EMS data provided spatial and temporal precision, hospital records added medically verified injury severity, and medical examiner data filled gaps in fatality reporting. The goal was not only to improve severity coding but also to achieve completeness, ensuring that every crash-related injury within the city was documented in the system. Soltani et al. (2022) used probabilistic matching to integrate over 17,000 injury records, finding that only about one-quarter appeared in both police and hospital datasets. Vulnerable groups, including low-income residents, pedestrians, and cyclists, were the most underrepresented in police data. Hosseinzadeh et al. (2022) also developed a heuristic framework in Kentucky to link police, EMS, and trauma registry data, achieving match rates above 70% between EMS and crash reports. The study identified thousands of additional injuries absent from police records, particularly non-fatal cases and those treated on-scene or with delayed hospital admission

A formal systematic review by Soltani et al. (2024) confirmed that such linkages consistently reveal significant gaps in police-reported crash data. Across studies from 1994 onward, hospital data reported significantly more injuries than police data, particularly for VRUs and less severe injuries. The review also found that comparing severity assessments across police, EMS, and trauma registries can help identify biases in police-reported data regarding injury severity classification. The systematic review also identified methodological barriers, including missing identifiers, privacy concerns, and inconsistent injury definitions, that constrain data integration efforts.

These review findings are consistent with earlier work by Flannagan et al. (2021) that provided a framework for a comprehensive approach to traffic crash injury measurement and reporting system. The report noted that pilot efforts in California, Kentucky, and Utah demonstrated that adding EMS and hospital data increases the number of identified serious injuries by 30 to 60 percent compared with police data alone.

International research corroborates the U.S. experience, demonstrating that underreporting is a systemic issue across law-enforcement data sources. In Australia, Watson et al. (2015) found that nearly two-thirds of hospital-treated crash injuries lacked a corresponding police report. The missing cases had a disproportionately higher rate of cyclists and single-vehicle crashes. In the Netherlands, researchers at the Institute for Road Safety Research (SWOV) found that police data could no longer serve as the primary source for estimating serious injury incidence. Weijermars et al. (2016) redefined "serious road injuries" based on hospital records and linked them to police data. They noted that a hospital-anchored approach results in substantially higher and more accurate injury counts. Loo and Tsui (2010) chose to use the hospital-based Road Casualty Information Database to analyze bicycle crashes in Hong Kong and cited the potential bias and incompleteness of the police-reported crash data in the urban contexts.

In Denmark, Janstrup et al. (2016) used capture–recapture models combining police and hospital data to estimate unobserved injuries. Their analysis revealed that police data captured fewer than two-thirds of total hospital-recorded injuries, with the largest discrepancies among cyclists. French studies from the 2000s (Amoros et al. 2006) also echoed that underreporting was most pronounced for low-severity and single-vehicle crashes, indicating structural biases in police-reported data.

## 2.3 Conclusions from Literature Review

In both U.S. and international contexts, the evidence consistently demonstrates that police crash databases alone fail to capture the true extent and distribution of road traffic injuries. This undercounting disproportionately affects vulnerable road users, distorting the spatial and demographic representation of road risk. In other words, the police data's incompleteness may be a systemic limitation of enforcement-based collision databases leading to several issues, such as the underestimation of the scale of road trauma and the misrepresentation of the spatial distribution of collisions as potentially the relative burden among different user groups. As Vision Zero policies rely on accurate injury mapping and performance monitoring, incomplete data, therefore, pose a structural barrier to achieving these goals. Using heuristic-based linkage with structured EMS data that includes fields such as incident date and GPS location can improve match accuracy, and the latter can also help in geolocating unmatched hospital records, anchoring them to roadway networks for geospatial analysis. However, use of these data at scale requires widespread adoption of standardized databases such as NEMSIS (National EMS Information System for Emergency Medical Services) ("National EMS Information System (NEMSIS)," |FHWA n.d.) and NTDS (ACS n.d.). Given the current state of these data—for example, NEMSIS data from only 2008 through 2012 are available as a public release dataset—we are far from replacing police-reported data as the main source for estimating high-injury networks in California communities. These practices of sourcing collision data from multiple sources and merging them remain resource-intensive and complicated by privacy regulations. While they are worth exploring for the estimation of HINs in the long run, in the short to medium term, we need more creative ways to complete the crash data.

# 3. Data Analysis: EMS and Police-Reported Data

To address the goal of expanding the collision data used to estimate high-injury networks, the first step was to use EMS data and verify whether the discrepancy is as common in the context of California as the literature suggests. Toward that end, the research team collaborated with the stakeholder/partner, the San Luis Obispo Council of Governments, which provided the reported traffic crash data from EMS collected as part of their Regional Safety Action Planning grant. We first used EMS collision data collected for the city of San Luis Obispo and compared them with what is reported for the same jurisdiction in the official SWITRS (Statewide Integrated Traffic Records System) data available at TIMS ("TIMS - Transportation Injury Mapping System" n.d.). This step helped us validate the literature findings in the context of regional data from California. It also helped us document the challenges of connecting the two sources of roadway injury data.

## 3.1 Context: City of San Luis Obispo

San Luis Obispo (SLO) is a city on California's Central Coast and the county seat, situated roughly midway between Los Angeles and San Francisco. The community is centered on Mission San Luis Obispo de Tolosa ("San Luis Obispo de Tolosa," California Missions Foundation n.d.) and has a major university (Cal Poly) with enrollment of more than 22,000 students ("Facts and Figures," Cal Poly n.d.). The city offers a high quality of life with extensive civic services and growing active-transportation amenities ("Neighborhood Traffic Management," City of San Luis Obispo, CA, n.d.). The mission is a cultural anchor for the downtown district (California Missions Foundation n.d.). Demographically, SLO is a mid-sized city of about 45,000 people within a county with a population of about 280,000 residents ("Living in San Luis Obispo," City of San Luis Obispo, CA n.d.). The local climate is Mediterranean, with mild temperatures that facilitate year-round outdoor travel and recreation. The city's climate, combined with its urban form characterized by short trip lengths and multiple neighborhood centers ("U.S. Census Bureau QuickFacts" n.d.), makes it conducive to walking and bicycling.

Based on the research team's knowledge, the city was divided into the following analytical zones, aligned with recognizable corridors and activity centers of the community. It allowed us to examine patterns between crashes reported in the EMS data and those found in the official SWITRS data (available via TIMS):

- Cal Poly Campus: This area is characterized by predominantly student travel along on-campus pathways and the surrounding perimeter streets, such as Grand Avenue and Perimeter Road. Pedestrian and bicycle activity is especially concentrated along the campus edges during class-change periods. The university's size and enrollment significantly influence seasonal travel demand patterns throughout the city.

- Downtown Core: The walkable grid surrounding Higuera, Marsh, Monterey, and Chorro Streets near the Mission supports dense retail and restaurant activity. Frequent street crossings and active evening and weekend periods contribute to consistently high pedestrian volumes in this area.

- Foothill/Highland: These residential connectors between the campus and adjacent neighborhoods, including Foothill Boulevard and California Boulevard, support a mix of local trips and student commuting. Travel volumes tend to increase in alignment with the academic calendar.

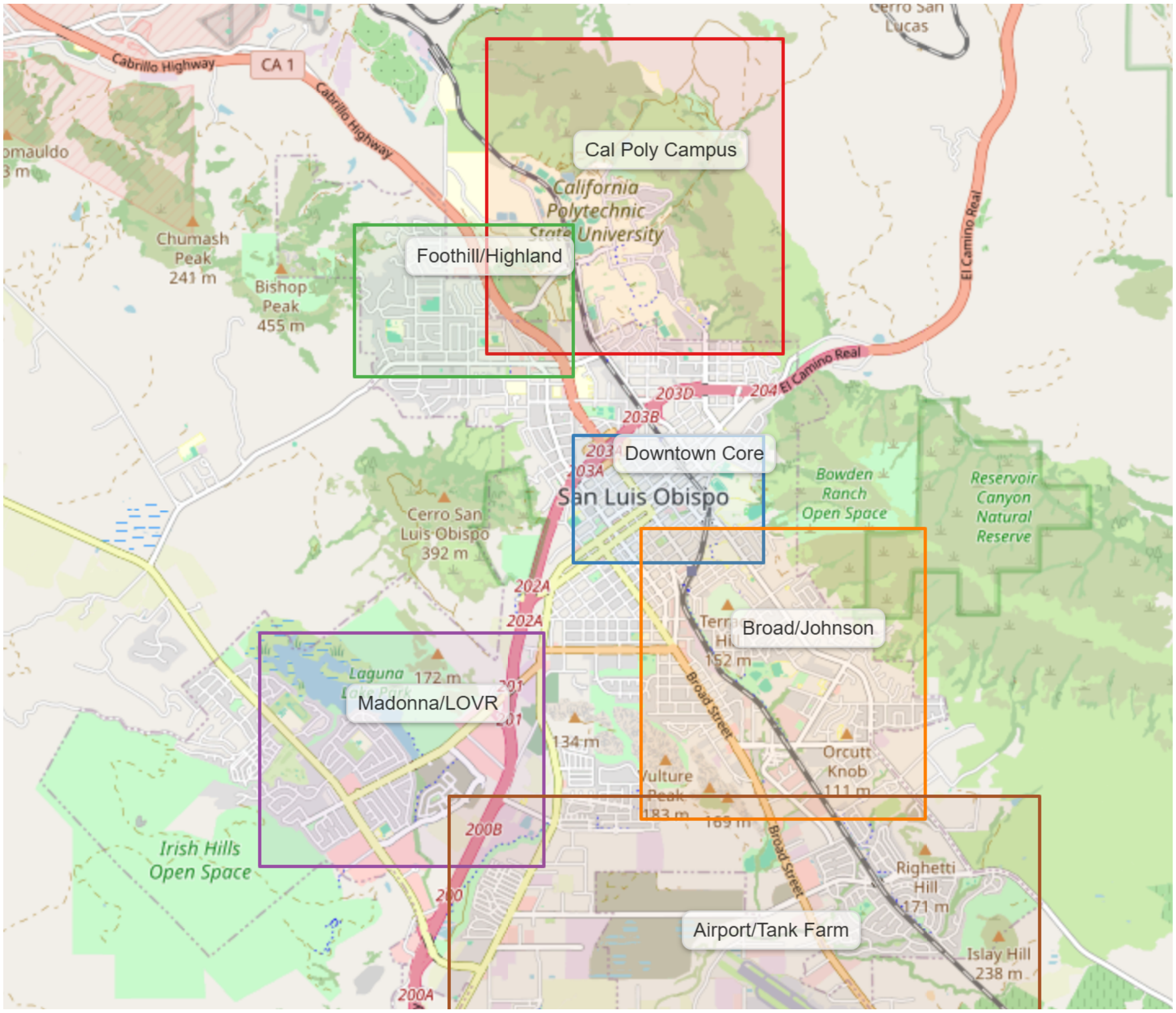
- Madonna/LOVR (Los Osos Valley Road): This area features wider commercial arterials that serve regional shopping destinations near Laguna Lake. Traffic volumes and speeds are generally higher here than on streets in the downtown area.

- Broad/Johnson: These neighborhood corridors link downtown to the Orcutt and Tank Farm areas are characterized by mixed residential frontage and moderate traffic speeds.

- Airport/Tank Farm: Located south and east of downtown, this area consists of business parks and wider roads near the regional airport and Tank Farm Road. The transportation network is comparatively more automobile-oriented than in other parts of the city.

These zones were used to compare where EMS incidents and TIMS police-reported crashes (primarily on public rights-of-way) diverge and to summarize “hot” and “cold” spots across datasets. These areas are shown in Figure 1.

Figure 1. Areas in the City of San Luis Obispo




## 3.2 Patterns in Data

To demonstrate the extent of differences in the collision data reported in San Luis Obispo by the two sources, we examined bicycle and pedestrian crash data for 2023. There were two sets of these crashes: one from EMS and one from TIMS. Both sets of these crashes were mapped onto the city of San Luis Obispo maps based on their reported locations. Figure 2 shows the EMS data heatmap, while Figure 3 shows the official TIMS data heatmap. Official TIMS data generally underreport collisions, but other patterns are also visible in the two figures. A summary of the patterns observed from the two sets of heatmaps is provided for each city area below.

Figure 2. Heatmap of the collisions from the EMS data in the City of San Luis Obispo

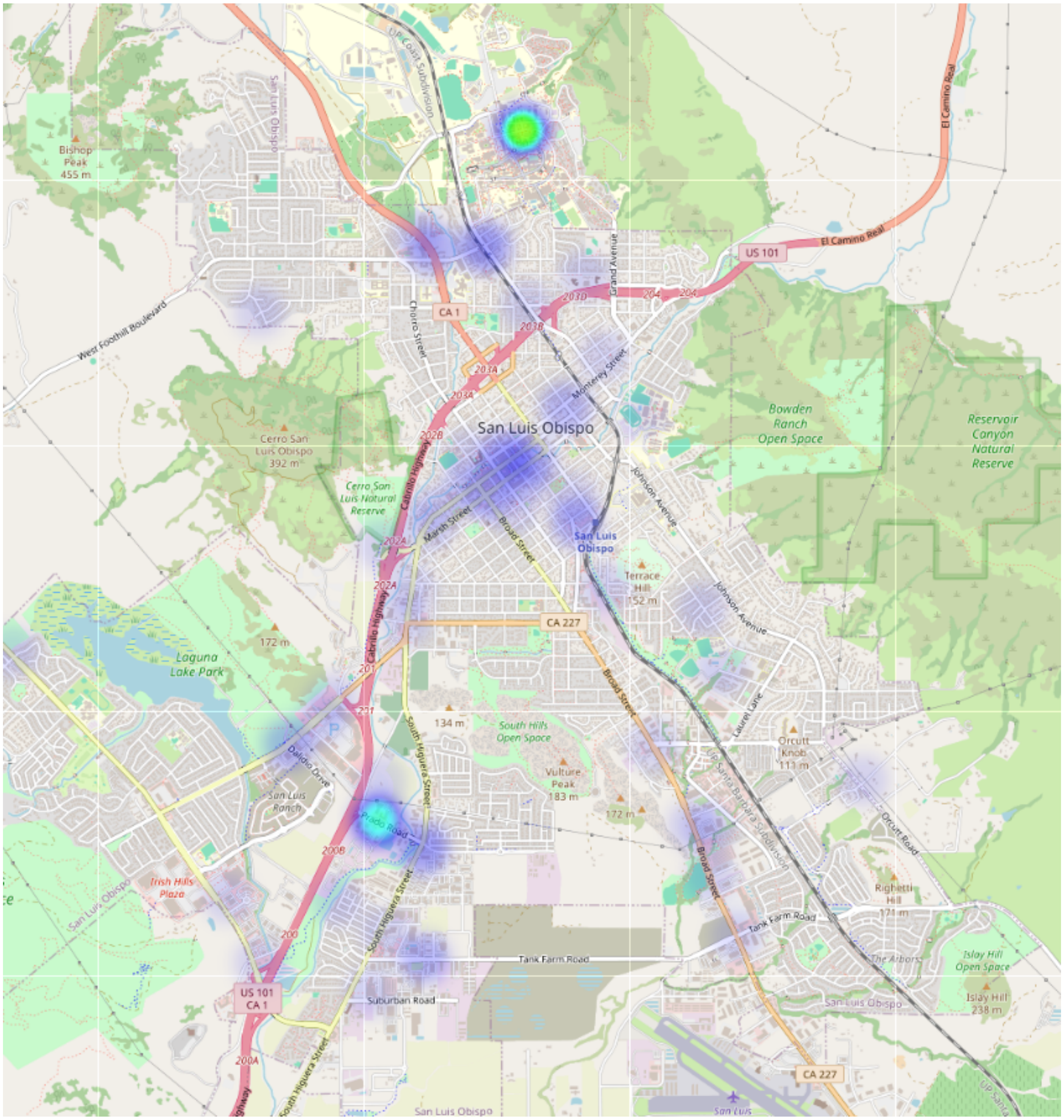
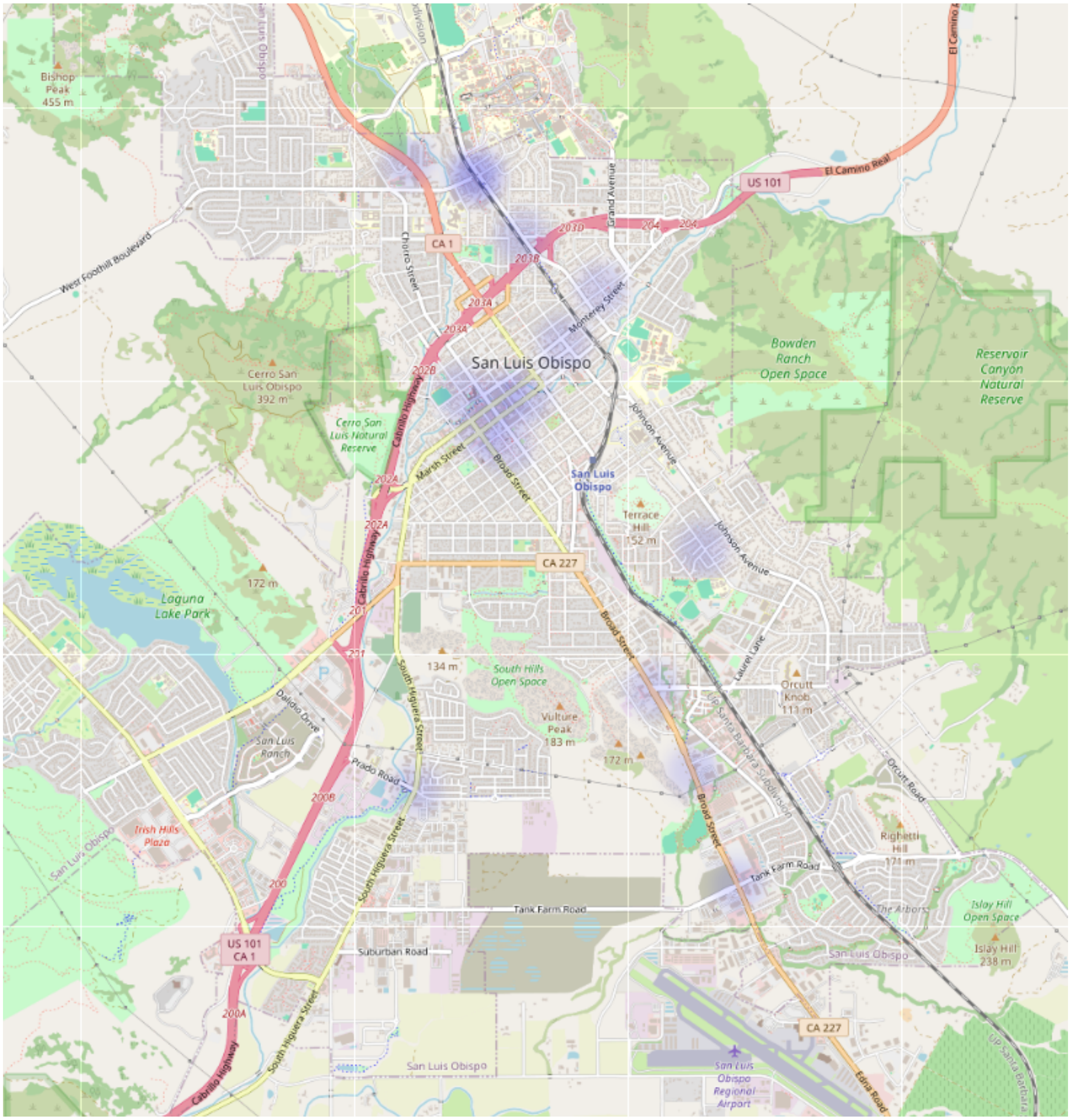


Figure 3. Heatmap of the collisions from the TIMS data in the City of San Luis Obispo




**Cal Poly Campus**

On and immediately adjacent to Cal Poly, along Grand Avenue, Perimeter Road, Poly Canyon, and Village Drive, the EMS heatmap is notably more pronounced than the TIMS heatmap. This pattern suggests that campus or campus-edge incidents often prompt EMS responses but do not always appear in police-reported crash data. Calls such as Grand Avenue at South Perimeter and Village Drive at Poly Canyon exemplify locations where EMS activity is generally higher than what is visible in the TIMS data.

#### Downtown Core

Within the downtown grid (i.e., Higuera, Marsh, Monterey, Morro, and Chorro streets), both datasets show activity. The emphasis here reflects the types of events likely to be reported as roadway crashes, including conflicts at urban intersections, mid-block interactions near parking turnover, and evening or weekend peaks around restaurants and shops. This is the area of the city where we find at least some crashes reported in the TIMS database but not in the EMS database.

#### Foothill/Highland – California Boulevard Corridor

The Foothill/Highland area, including the California Boulevard spine, appears in both datasets. As with downtown, TIMS maintains a footprint here as well. This corridor connects student neighborhoods and campus access points, producing a steady baseline of reportable roadway incidents alongside medical responses.

#### Madonna/LOVR Corridor

Along Madonna Road, Los Osos Valley Road, and Laguna Lake approaches, TIMS again outweighs EMS. The wider, faster arterials and retail driveways in this district are more likely to generate police-reported crashes, which the TIMS heatmap captures more fully than EMS.

#### Broad/Johnson – Orcutt – Tank Farm

In the Broad and Johnson corridors extending toward Orcutt and Tank Farm, TIMS shows consistent coverage, while EMS registers fewer direct address matches. The roadway context of collector and minor arterial streets with mixed residential and employment land uses aligns with incidents more frequently documented in crash reports than in medical-only calls.

The first map (EMS) highlights areas on or immediately adjacent to Cal Poly (e.g., Grand Ave/Perimeter/Poly Canyon/Village Dr) that don't show up in the second map (TIMS). Conversely, the TIMS heatmap emphasizes arterials and downtown streets (e.g., Higuera/Marsh/Monterey, Madonna/LOVR, Broad/Johnson) where police-reported roadway crashes are more common.

## 3.3 Conclusions

The analysis presented herein clearly shows that EMS and SWITRS data share some cases, but several crashes (including those involving bicyclist and pedestrian injuries) appear in the EMS data but not in the SWITRS data collected via TIMS. It clearly demonstrates that high-injury networks generated from SWITRS data do not capture the full picture of the traffic safety issues faced by a community. Through email communication with SLO County EMS, the researchers learned that comparing the two is extremely difficult due to their contrasting fields of expertise. Law enforcement lacks EMS training beyond basic first aid/CPR, and EMS lacks law enforcement training. These differences are expected, as both sets of professionals have completely different objectives (public safety and traffic control vs. medical care). It is relevant to note that, for any reported incident, having law enforcement on scene does not necessarily mean EMS will be there, and vice versa. SLO County EMS staff also reiterated what the authors found in the literature regarding patients' injury severity: EMS data likely provides an accurate assessment of it.

Overall, the findings are consistent with the literature about the incompleteness of the collision data. If one wants to capture a complete picture of reported incidents, in the long run, ensuring better coordination between law enforcement and EMS personnel should be the way forward. In the short term, we need to explore an alternative data set. The work by Calma and Jackson (2021) used 911 call data to identify collision patterns in Washington, D.C., and in the next chapter, we expand on the use of 911 call data and assess the differences between 911 call data and police-reported collisions.

# 4. Data Analysis: San Francisco 911 Call Data

Based on an analysis of EMS data compared with police-reported SWRITS data collected via TIMS, neither source provides a complete picture of collisions in a community. The 911 call data were a potential source to gather a complete picture of all traffic-related collisions in a community. This chapter provides details of scraping and analyzing the traffic-related incidents reported on the 911 call logs from PulsePoint (PulsePoint n.d.). The work was carried out using data from the city of San Francisco. The city of San Francisco was selected since the official crash records are made available by the city within a short interval after their occurrence (about a 45-day lag) ("Map of Traffic Crashes Resulting in Injuries | DataSF" n.d.). Furthermore, given that San Francisco is a larger and more diverse community, it allowed us to explore whether police-reported data show biases based on the socioeconomic characteristics of the crash location. In this chapter, we first describe the context of San Francisco, then the process for collecting and archiving 911 call data, and finally the results of the analysis.

## 4.1 Context: City of San Francisco

San Francisco is a compact, 47-square-mile peninsula divided into census tracts and Public Use Microdata Areas (PUMAs). The City's Planning Department also aggregates tracts into "Analysis Neighborhoods" to publish small-area social and economic indicators, a practical stand-in for "census zones" ("Index of /Geo/Maps/DC2020/PUMA/St06_ca" n.d.).

Post-2020, population and migration patterns have resulted in a modest net decline in population, with the current population in the mid-800,000s range, skewing relatively younger and more highly educated for a U.S. city ("San Francisco, CA | Data USA" n.d.). The median household annual income for the city is above $140,000, with significant Asian and white non-Hispanic populations and smaller but historically important Black and Latino communities. From a socioeconomic perspective, there is stark inequality across geographic areas: high-income cores around the northeast waterfront and Noe Valley-west ridges contrast with lower-income pockets in parts of the southeast (Bayview/Visitacion Valley) and the Tenderloin. These patterns are visible in tract-level health and poverty maps ("San Francisco, CA | Data USA" n.d.). The political economy is anchored by tech, biotech, tourism/conventions, education/health care, and port-adjacent logistics. Downtown's office market and transit ridership are still normalizing post-pandemic, yet city economic reports show labor-market resilience alongside tech contraction; public sector and hospitality have helped offset losses, while recovery strategies focus on mixed-use downtown and neighborhood commercial corridors ("Snapshot of the SF Economy As of May 2025" n.d.).

Eastern parts of the city (SoMa, the Financial District, Mission Bay) concentrate jobs, higher-rent multifamily housing, and transit; southeastern tracts (Bayview/Hunters Point, Vis Valley) show more families, industrial legacy parcels, and environmental justice priorities; western tracts (Richmond, the Sunset, Parkside) are more owner-occupied, with mid-density blocks and strong ethnic enclaves; central ridge tracts (Noe Valley/Haight/Castro) blend high incomes, older housing stock, and transit access. These patterns guide resource allocation, health equity programming, and capital planning published by the City ("Analysis Neighborhoods | DataSF" n.d.).

San Francisco Police Department (SFPD) districts (for everyday governance and statistics) roughly align with neighborhood clusters, and the City maintains a current boundary map and station locator ("Map of Current Police Districts | DataSF" n.d.). A brief description of each police district is provided below:

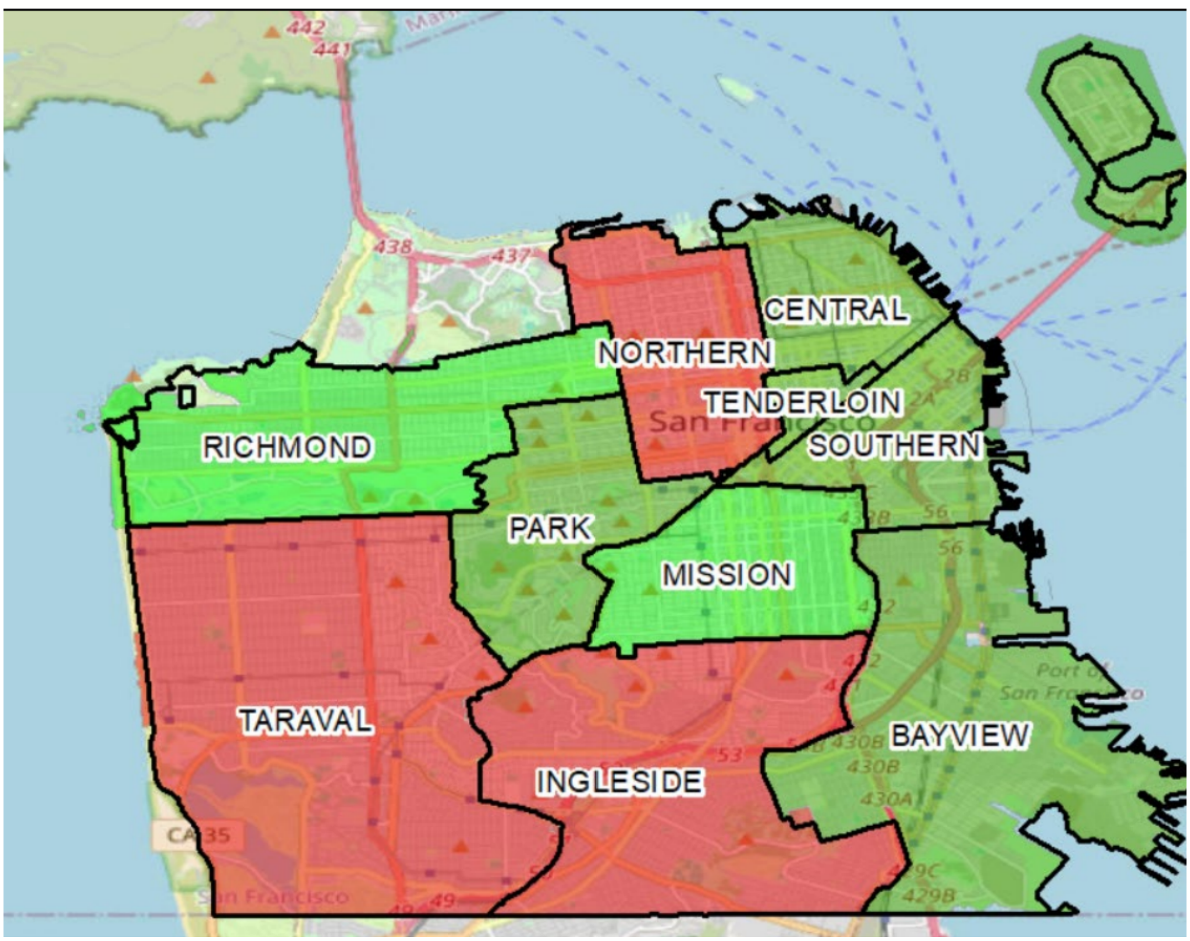
- Central: Embraces Chinatown/North Beach/Waterfront; dense tourism economy, heavy pedestrian flows.
- Southern: SoMa/Mission Bay; major event venues and new residential towers; peak property-crime pressure in recent cycles.
- Tenderloin: High-need services cluster; intensive street-level operations and social-service coordination.
- Northern: Marina/Cow Hollow/Russian Hill; mixed entertainment strips and affluent residential hills.
- Mission: Mission/Noe Valley/Dolores; vibrant commercial corridors and nightlife; long-standing equity and displacement focus.
- Bayview: Bayview/Hunters Point/Visitacion Valley; industrial waterfront transition, environmental justice investments.
- Park: Haight-Ashbury/Golden Gate Park; destination recreation areas and visitor management.
- Richmond: Inner/Outer Richmond; diverse immigrant communities, neighborhood retail, and coastal parks.
- Taraval: The Sunset/Parkside/Lakeshore; largely residential, family-oriented west-side tracts with heavy fog microclimate.
- Ingleside: Excelsior/Ocean View/Balboa Park; working- and middle-class southeast hills with key transit hubs.

Figure 4 shows the district boundaries overlaid on the city map. The estimated population for each police district and its share of the population are reported in Table 1.

Table 1. San Francisco Police Districts and their Population and Share of the City Population

| Police District | Population | % of Total Population |
|---|---|---|
| Bayview | 73,864 | 8% |
| Central | 69,929 | 8% |
| Ingleside | 136,564 | 16% |
| Mission | 81,785 | 9% |
| Northern | 103,787 | 12% |
| Park | 63,325 | 7% |
| Richmond | 87,094 | 10% |
| Southern | 65,115 | 7% |
| Taraval | 153,038 | 18% |
| Tenderloin | 35,841 | 4% |

Figure 4. Map of San Francisco Police Districts



Together, these census-based zones and police districts reveal a city where small distances yield substantial differences in community characteristics. We used this context to analyze traffic collision patterns suggested by an analysis of the data from 911 calls (PulsePoint n.d.) as well as from the official city-reported data.

## 4.2 Collision and Incident Data Collection

Police-reported crashes for the city of San Francisco were collected using DataSF ("Map of Traffic Crashes Resulting in Injuries | DataSF" n.d.). Traffic injury crash data are added by the city approximately once a month, with a relatively small lag (about 45 days), so data from two months ago are reliably available at any given time. The variables associated with each collision include the time and location of the incident, the severity of injury to the involved parties, and the type of collision. While the team is still collecting data from PulsePoint for future analysis, this report only analyzed the data available from the City of SFO through 9/15/2025. These official police-reported data are compared with 911 call data collected from PulsePoint. The data used here are from 4/16/2025 through 7/31/2025. Note that the city has data going back several years, but in this research, we are limited by the scraped 911 call data, and the research team was only able to get that fully set up starting 4/16/2025, since PulsePoint, the source of 911 call data, doesn't archive the data, and it has to be collected in real-time.

PulsePoint is integrated into call centers and provides real-time data to the public. Many types of emergencies, such as fires, medical emergencies, and traffic collisions, can be found on PulsePoint (PulsePoint n.d.). To ensure all traffic collision data were recorded, a scraping program was used. The scraper automatically archives 911 call data, eliminating the need for anyone to monitor PulsePoint continuously. Our PulsePoint scraper was provided by the team from "Data Driven Streets," who have implemented this for several communities throughout the U.S. The scraper runs every 10 minutes, which is frequent enough to ensure all PulsePoint calls get logged.

The PulsePoint scraper collects data for all types of emergencies, and for this study, traffic collisions were the incidents of interest. These are listed as "TC" and "TCE" under the "pulsepointincidenttype" field. TC stands for "traffic collision," and TCE stands for "traffic collision expanded." Typical expanded traffic collision features include higher speeds, multiple parties involved in the incident, overturned vehicles, and serious injuries. Collisions that do not have any of these more extreme elements are classified as "TC" (PulsePoint n.d.).

## 4.3 Data Cleaning and Processing

To match city-reported data from DataSF with the PulsePoint 911 call data, some processing was required. For example, the incident time was provided in UTC (Coordinated Universal Time) and it was converted to Pacific Time. Both time variables were then converted to the same Date-Time format.

To determine the extent to which traffic collisions reported in the 911 call data are reflected in the city's official databases, city-reported incidents need to be matched with PulsePoint incidents. This was accomplished using the time, location, and street name parameters common to both datasets. Matching was performed in ArcGIS, which enables the geographic visualization of collisions. The Excel file with incidents was loaded into ArcGIS, and the collisions were projected above a map of San Francisco. To ease visual matching, points were colored according to the source column. Incidents were filtered one day at a time. Once all matches were found for the day, we filtered on the next day. If an official city database reported a traffic collision, and a PulsePoint traffic collision occurred at a similar location and time, we recorded them as the same incident. This matching process was somewhat subjective, as the city's reported time and location of the collision rarely exactly matched the 911 call time, but almost all incidents recorded as a pair had a reported time within 10 minutes. Locations were typically very close, within a few feet of each other. There were some records (less than 4%), however, that were as far as 500 ft. Once this matching process was completed across the entire dataset, data from both sources (DataSF and PulsePoint) were combined into a single spreadsheet.

Variables that matched across both sources, such as location coordinates, time, and street names, were in the same column. Variables that were unique to only one dataset were kept in separate columns. A source column was added to know which source each observation came from. As expected, based on the literature review and the premise of this work, not all traffic collisions reported on PulsePoint have a matching incident in the police-reported database, and vice versa. Note that there were substantially fewer cases of police-reported collisions not reported on PulsePoint. Based on the reported incident locations, a GIS file from the U.S. Census Bureau was used to identify the census tracts for each incident. The census tract was used to determine the demographics of the area where the incident occurred and to identify the corresponding police district.

## 4.4 Overlap Between PulsePoint and Official City Crash Data

The next step was to pull only the PulsePoint-sourced records from the combined dataset of traffic incident records from both PulsePoint and the City, and to create several views of the data to better understand how these incidents are distributed across San Francisco. For each PulsePoint data point, a binary variable indicated whether its match was found in the corresponding collision data from the city. To demonstrate their spatial differences, we focused first on the census tract level. For each census tract, we calculated the proportion of incidents that were unpaired (pair = 0) versus paired (pair = 1) and compared how these proportions varied across the city. However, in some census tracts, we had very few incidents, which skewed the proportions. To ensure we had a unit of analysis with a meaningful number of observations per unit, we aggregated the data to the police district level to obtain a neighborhood-scale view. This included counting the number of PulsePoint incidents in each police district and calculating the proportions of incidents with the variable pair = 0 and pair = 1.

Based on that, we could estimate the proportion of PulsePoint incidents that could be matched to a corresponding case in the police data and those that could not. The results of the analysis by police districts are shown below.

Table 2. Proportion of Matched Cases by Police District

| SFPD Police District | Proportion of unmatched incidents (pair = 0) | Proportion of matched incidents (pair = 1) | Difference |
|---|---|---|---|
| Mission | 0.40 | 0.60 | -0.20 |
| Central | 0.43 | 0.57 | -0.14 |
| Northern | 0.43 | 0.57 | -0.13 |
| Park | 0.49 | 0.51 | -0.02 |
| Tenderloin | 0.49 | 0.51 | -0.02 |
| Taraval | 0.52 | 0.48 | 0.04 |
| Richmond | 0.56 | 0.44 | 0.11 |
| Ingleside | 0.57 | 0.43 | 0.14 |
| Southern | 0.58 | 0.42 | 0.17 |
| Bayview | 0.59 | 0.41 | 0.18 |

Table 2 shows the proportion of PulsePoint incidents with official city records to assess how consistently PulsePoint events are captured in the official database across San Francisco police districts. The districts in the table above are sorted based on the difference between the proportion of unmatched PulsePoint incidents (pair = 0) and matched incidents (pair = 1). Districts with negative differences show stronger alignment with official data (higher matching), while positive differences indicate a larger share of PulsePoint incidents that do not appear in official records.

The results show clear geographic variation. High-matching districts, where most PulsePoint incidents are reflected in official databases, include the Mission, Central, Northern, and the Tenderloin. Medium-matching districts, showing a more balanced mix of matched and unmatched incidents, are Park and Taraval. Low-matching districts, where PulsePoint captures substantially more incidents than appear in official records, include Bayview, Southern, Ingleside, and Richmond. PulsePoint data likely adds the most additional information in low-matching districts, while in high-matching districts, it essentially reinforces patterns visible in official data. The four low-matching districts are moderately dense residential areas where PulsePoint captures substantially more incident activity than is reflected in official databases. In contrast, the highest-matching districts are denser and more commercially active, suggesting that this increases consistency between community-reported and official incident records. This suggests that land use, reporting practices, and institutional presence may play a larger role than population density alone in explaining differences in matching.

This analysis was followed by a similar analysis using only the incidents in the PulsePoint data marked as “TCE,” which denotes “Traffic Collisions Expanded.” As mentioned previously, these incidents featured higher speeds, multiple parties, overturned vehicles, and serious injuries. As expected, a higher proportion of these incidents from the PulsePoint data made it to the official database. The difference, however, is not substantial (48.1% vs. 50.3%) (see Table 3).

Table 3. Proportion of Matched Cases for Overall PulsePoint Data vs. TCE only Incidents

| Proportion | Full PulsePoint data | PulsePoint data (TCE-only) |
|---|---|---|
| Unmatched incidents | 0.519 | 0.497 |
| Matched incidents | 0.481 | 0.503 |

The results of the analysis of TCE-only incidents by police districts are shown in Table 4.

Table 4. Proportion of Matched Cases by Police Districts (TCE only)

| Rank | SFPD Police District | Proportion of unmatched incidents (pair 0) | Proportion of matched incidents (pair = 1) | Difference (0–1) |
|---|---|---|---|---|
| 1 | The Mission | 0.34 | 0.66 | −0.319 |
| 2 | Northern | 0.41 | 0.59 | −0.184 |
| 3 | The Tenderloin | 0.45 | 0.55 | −0.100 |
| 4 | Central | 0.48 | 0.52 | −0.037 |
| 5 | Bayview | 0.52 | 0.48 | +0.032 |
| 6 | Park | 0.53 | 0.47 | +0.067 |
| 7 | Southern | 0.54 | 0.46 | +0.077 |
| 8 | Taraval | 0.56 | 0.44 | +0.125 |
| 9 | Ingleside | 0.57 | 0.43 | +0.130 |
| 10 | Richmond | 0.62 | 0.38 | +0.231 |

For TC-expanded (TCE) calls, these geographic patterns become more pronounced. The Central, Mission, Northern, and Tenderloin districts demonstrate high matching rates, with a strong majority of TCE PulsePoint incidents matching official records. Conversely, the Richmond, Ingleside, Taraval, and Southern districts show markedly lower matching rates, indicating that a substantial portion of TCE incidents captured by PulsePoint do not appear in official databases.

Table 5 compares results from the overall database with those from TCE-only cases.

Table 5. Percentage of Matched Cases: A Comparison Between TC and TCE Incidents

| Police District | Full PulsePoint: % Matched | TCE-Only: % Matched | Interpretation |
|---|---|---|---|
| Mission | 59.9% | 65.9% | Highest matching in both datasets |
| Central | 57.0% | 51.9% | Strong and consistent matching |
| Northern | 56.7% | 59.2% | High alignment with official data |
| Tenderloin | 50.8% | 55.0% | High matching, especially for TCE |
| Park | 51.2% | 46.7% | Mixed but relatively balanced |
| Taraval | 48.1% | 43.8% | Lower matching for TCE calls |
| Bayview | 40.8% | 48.4% | Low overall matching, but higher matching for TC Expanded |
| Southern | 41.7% | 46.2% | Low matching in both datasets |
| Ingleside | 42.9% | 43.5% | Low matching in both datasets |
| Richmond | 44.3% | 38.5% | Lowest alignment, especially for TCE |

In both the full PulsePoint dataset and the TCE-only subset, Central, the Mission, Northern, and the Tenderloin districts exhibit the highest matching rates, indicating that most PulsePoint incidents in these areas are also captured in official databases. In contrast, western and southern districts (i.e., Richmond, Ingleside, Taraval, Southern, and Bayview) show substantially lower matching rates, meaning PulsePoint records a larger share of incidents that do not appear in official data. These patterns suggest that PulsePoint provides its greatest additional informational value in more residential districts, while in denser, downtown-oriented districts, it largely mirrors existing official reporting.

## 4.5 Stepwise Regression Modeling

Stepwise regression models were estimated to test relevant hypotheses regarding 911 call data and/or police-reported data from the city. The first hypothesis was whether, within the 911 call data from PulsePoint, TC (not extended) and TCE (extended incident) incidents differ statistically and, if so, which variables associated with the location of those incidents best explain the difference between the two sets of data. To answer this question, we conducted a stepwise logistic regression with TC vs. TCE as the binary target variable. The various models were considered in an organized manner to find the best-fit model. The best-fit model had the lowest AIC (Akaike Information Criterion) and the lowest BIC (Bayesian Information Criterion). Details on these measures to assess model fit may be found in Ben and Yohai (2004). The following variables were tested as explanatory variables:

- Incident location (defined by latitude and longitude)
- Incident location (public vs. private)
- Incident location (census tract)

- Population of the census tract
- Population density of the census tract (population/km$^2$)
- Median household income
- Percentage of minority households
- Percentage of households with a member with a bachelor's degree or above in the census tract
- Time category defined based on time of day and day of week for the incident (morning peak period; afternoon peak period; Friday/Saturday night; other off-peak period)

The model fit criteria for each tested model are shown in Table 6. The last row of the table shows the best-fit model estimated using the stepwise approach. The best-fit model with the lowest values for AIC and BIC had the following predictors:

- Total population
- population density/km$^2$
- median household income

while controlling for public location (binary variable) as well as for whether the incident had a corresponding match in the police data (binary variable).

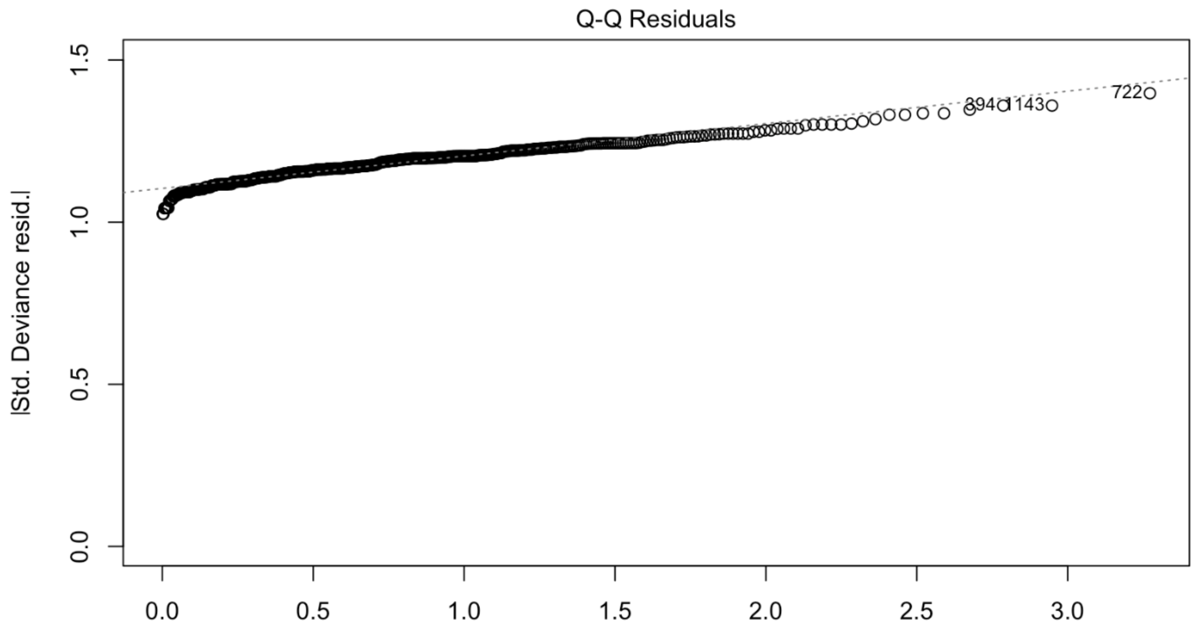
The Q-Q plot of the residuals from the best-fit model is shown in Figure 5 below. This plot shows that among the 911 call data, the difference between TC and TCE can be explained by the best-fit model.

Table 6. AIC and BIC Values Associated with the Logistic Regression Models for TC vs. TCE Incidents in the PulsePoint Data

| Model | Explanatory Variables | AIC | BIC |
|---|---|---|---|
| 1 | Location coordinates + time-category | 1539.2 | 1569.6 |
| 2 | Location coordinates + time category controlled for public location | 1161.3 | 1189.8 |
| 3 | Census tract controlled for public location | 1301.8 | 2215.1 |
| 4 | Total population + population density/km2 controlled for public location | 1144.9 | 1159.1 |
| 5 | Total population + population density/km2 + median household income controlled for public location | 1111.6 | 1130.4 |
| 6 | Total population + population density/km2 + median household income + race % information controlled for public location | 1113.7 | 1151.5 |
| 7 | Total population + population density/km2 + median household income + education information (percentage with graduate or bachelor's degree) controlled for public location | 1112.7 | 1141 |
| 8 | Total population + population density/km2 + median household income + school education information controlled for public location (high school %, less than high school %) | 1113.1 | 1141.3 |
| 9 | Total population, population density/km2, median household income controlled for whether police reports and 911 call were matched | 742.45 | 759.57 |
| 10 | Total population, population density/km2, median household income controlled for whether police reports and 911 call were matched (binary) as well as public location (binary) | 655.85 | 672.45 |

The Q-Q plot of the deviance residuals from the best-fit model is given below. The deviance residual plot is used for GLMs as opposed to the usual Q-Q plot for data that is normally distributed (Ben and Yohai 2004). This plot shows that among the 911 call data from PulsePoint, the difference between TC and TCE can be well explained by the best-fit model discussed.

Figure 5. Q-Q Diagnostic Plot for the TC vs. TCE Incidents Logistic Regression Model in the PulsePoint Data




The final model and its logistic regression coefficients are given in the Table 7. The coefficients are not statistically significant at a 90% Confidence Interval, but the model explains the response variable (i.e., the proportion of TC among TC+TCE data) best among the stepwise models tested. The lack of significance may be due to multicollinearity, i.e., the effects of the explanatory variables being diluted by correlations between variables.

Table 7. Logistic Regression Model for TC vs. TCE Incidents in the PulsePoint Data

| Coefficients | Estimate | Std. Error | Z value | p-value |
|---|---|---|---|---|
| (Intercept) | 4.881e-01 | 3.896e-01 | 1.253 | 0.210 |
| Total Population | 3.536e-06 | 7.266e-05 | 0.049 | 0.961 |
| Pop. Density/km2 | -1.040e-05 | 1.609e-05 | -0.646 | 0.518 |
| Median Household Income | -2.486e-06 | 2.063e-06 | -1.205 | 0.228 |

The second hypothesis tested was whether, among all incidents, the data for which 911 calls and police reports matched statistically differed from those for which they did not and which demographic and incident-specific variables best explain this difference. The AIC and BIC values for each of the models considered are shown in Table 8, and, as noted previously, the best-fit model is the one with the lowest AIC and BIC values. The best-fit model included the following significant predictors: total population, population density/km$^2$, call type (TC, TCE, and other), time category (including time of day and day of week), and median household income, controlled for public location (binary variable).

Table 8. AIC and BIC Values Associated with Logistic Regression Models for Matched (y = 1) vs. Unmatched Incidents (y = 0)

| Model | Explanatory Variables | AIC | BIC |
|---|---|---|---|
| 1 | Source (police reports or 911 calls) | 2190.5 | 2201.5 |
| 2 | Model 1 + spatial coordinates | 2192.3 | 2214.4 |
| 3 | Call type (TC, TCE, missing (i.e., record only in the official city data[A1] [A2] )) | 1482.2 | 1502.4 |
| 4 | Call type controlled for public location | 1078.3 | 1097.3 |
| 5 | Time category (time of day, day of week) | 2445.9 | 2467.9 |
| 6 | Model 5 controlled for public location | 1177.1 | 1196.2 |
| 7 | Call type, time-category controlled for public location | 1078.8 | 1112.1 |
| 8 | Call type, time-category, population density/km2, total population controlled for public location | 1053.4 | 1096.1 |
| 9 | Model 8 + spatial coordinates, controlled for public location | 1056.8 | 1109 |
| 10 | Call type, time-category, **population density/km2, total population, median household income, controlled for public location (binary variable). | 1022.9 | 1070 |
| 11 | Population density/km2 controlled for public location (binary variable) | 1147.7 | 1157.2 |
| 12 | Call type, population density/km2 controlled for public location (binary variable) | 1053.6 | 1096.1 |

[A1]Please clarify.
[A2]Clarified since no call type is noted for records that were only in the official city reported data.

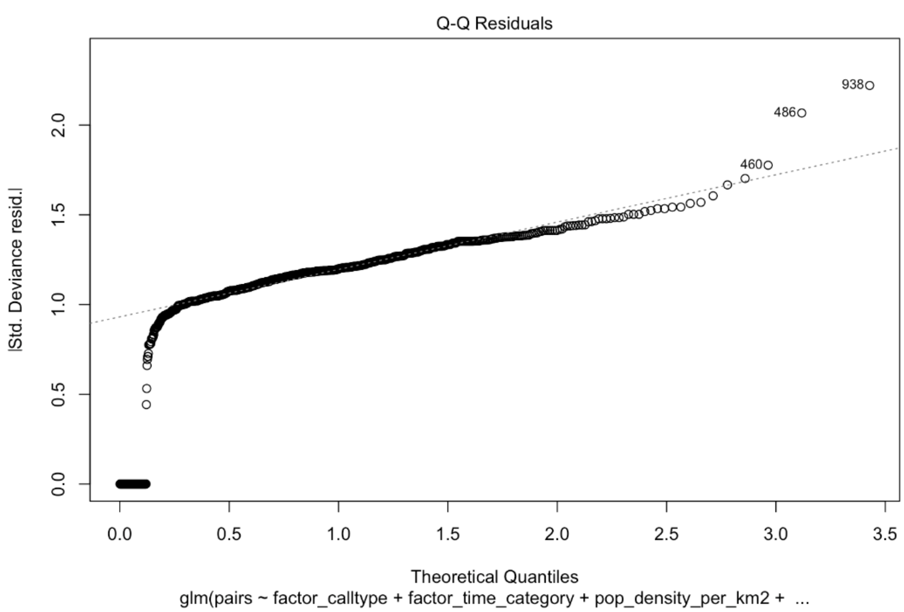
### 4.5.1 Discussion of Stepwise Regression Models

The final logistic regression model and its coefficients are given in Table 9. The model controls for the public location variable = 1, i.e., discards data reported at a non-public location. None of the coefficients are significant, except population density/km$^2$, but the model explains the response variable best among the stepwise models tested. According to the model, incidents that happen in high-density areas are more likely to be matched. The lack of significance may be due to multicollinearity, i.e., the effects of the explanatory variables being diluted. The time categories (weekday afternoon peak, weekday morning peak, Friday/Saturday night) were used as a categorical variable, with the base case set to other off-peak periods. Similarly, the call type (other, TCE, TC) was used as a categorical variable. Compared to other off-peak periods, weekday afternoon incidents are more likely to be matched, while Friday/Saturday night incidents are least likely to be matched.

Table 9. Logistic Regression Model for Matched vs. Unmatched Incidents in the PulsePoint Data

| Coefficients | Estimate | Std. Error | Z value | p-value |
|---|---|---|---|---|
| (Intercept) | 1.713e+01 | 1.973e+03 | 0.009 | 0.99307 |
| Weekday afternoon | 3.380e-01 | 2.114e-01 | 1.599 | 0.10976 |
| Weekday morning | 4.423e-02 | 2.658e-01 | 0.166 | 0.86784 |
| Weekend night | -6.878e-01 | 4.036e-01 | -1.704 | 0.08837 |
| Call type TC | -1.752e+01 | 1.973e+03 | -0.009 | 0.99292 |
| Call type TCE | -1.714e+01 | 1.973e+03 | -0.009 | 0.99307 |
| Population density/km2 | 4.266e-05 | 1.520e-05 | 2.806 | 0.00502** |
| Total population | 6.445e-05 | 5.823e-05 | 1.107 | 0.26836 |
| Median household income | -1.277e-06 | 1.662e-06 | -0.769 | 0.44217 |

Figure 6. Q-Q Diagnostic Plot for the TC vs. TCE Incidents Logistic Regression Model in the PulsePoint Data




The Q-Q plot of the deviance residuals is given for the “best” performing model. The diagnostic plot shows deviations at either end, suggesting the best-fit model may not hold well at the extremes.

## 4.6 Implications for High-Injury Network Development

The results clearly demonstrate that reliance on police-reported crash data alone can result in incomplete and uneven identification of injury risk, particularly for vulnerable road users (VRUs) and in certain roadway and neighborhood contexts. The policy guidance in this section provides actionable steps for transportation agencies to strengthen the development of HINs by incorporating supplemental injury data sources to address some challenges in the short- to medium-term.

First, enforcement-based data systems should be assumed to provide an incomplete picture of injuries in urban communities, as they often miss significant harm occurring on the transportation network. Police-reported crash data, therefore, while essential, should be treated as a foundation to be supplemented (but not replaced) by additional sources such as EMS records and 911 call data. Patterns of injury underreporting are not uniform and vary meaningfully by land use, roadway function, and location context, underscoring the importance of situational awareness in analysis.

These results highlight that the process of developing high-injury networks for urban areas should be updated to incorporate the regularized use of new and improved data streams. Toward that end, transportation agencies, especially those in the metropolitan regions where substantial portions of the street network are not managed by state DOTs, should formally incorporate Emergency Medical Services (EMS) data and 911 call records related to traffic collisions into the processes used to estimate HINs. This will not only help complete the data for injuries but also provide a more accurate picture of injury severity levels. The 911 call data can help complete the picture, while EMS data may be useful for accurately estimating injury severity.

The process documented here for scraping 911 call data would be a good starting point for agencies looking to initiate the integration. However, given the complexity and spatial patterns of the data, institutional capacity issues warrant consideration, and therefore, focusing initial efforts on pilot neighborhoods/corridors followed by phased expansion would be appropriate. For the pilot neighborhoods/corridors, both conventional HINs and supplemental injury networks derived from integrated data may be produced to document the value of using these data sources. As practices with additional data become more mature, incorporating multiple data sources may benefit from context-sensitive spatial weighting, given that the completeness of police-reported data varies across spatial and land-use contexts. Adopting this enhanced framework for the development of HINs will enable agencies to move closer to achieving Vision Zero.

# 5. Summary and Conclusions

## 5.1 Summary of Findings

This research examined the adequacy of police-reported crash data as the sole basis for identifying HINs with a particular focus on vulnerable road users (VRUs) in California communities. Through a literature review, empirical comparison of Emergency Medical Services (EMS) and police-reported data in San Luis Obispo, and a novel analysis of 911 call data from PulsePoint in San Francisco, the study demonstrates that traditional crash databases do not fully capture the spatial and contextual distribution of traffic injuries.

The analysis of EMS data from San Luis Obispo showed numerous pedestrian and bicycle injury incidents, particularly near the flagship university campus, that are absent from SWITRS police-reported records. Conversely, police data were more complete on major arterials and in the downtown area. These differences highlight how institutional reporting practices and roadway context shape official safety datasets and, hence, HINs. Neither EMS nor SWITRS data provided a full picture of collisions in the community, prompting the exploration of a novel data source, 911 calls from PulsePoint, to examine collision patterns.

Analysis of 911 calls from San Francisco demonstrated that 911 call data provide substantial additional information beyond official crash databases. A large share of traffic collisions reported via PulsePoint did not appear in police-reported records, with pronounced geographic variation across police districts. Lower-density, more residential districts exhibited lower matching rates, while denser and more institutional districts showed higher correspondence. Regression analyses indicated that population density, time of occurrence, and incident type influence whether a collision is recorded in official records, whereas neighborhood demographic variables were not statistically significant predictors.

Collectively, these findings support the conclusion that a reliance on police-reported data alone leads to incomplete and potentially inequitable identification of HINs. Integrating EMS and 911 call data offers a practical pathway to improve the accuracy, timeliness, and representativeness of safety analyses supporting Vision Zero implementation.

## 5.2 Limitations

Given that PulsePoint data are available only in real time and are not archived, this research initiated the scraping of 911 call data for San Francisco, which limited the temporal scope of the analysis. A smaller sample size resulting from this means that regression models may be affected by multicollinearity, reducing statistical power for some variables. The researchers have continued collecting and archiving the scraped data so that it is available for future analysis alongside a larger dataset.

Matching 911 calls to official crash records required judgment due to differences in timing and location reporting, which may introduce minor classification errors, even though these wouldn't affect the overall conclusions.

It is also noteworthy that 911 calls represent reported emergencies rather than verified crashes and, therefore, should be interpreted as a complementary data source and not as a replacement for SWITRS data.

## 5.3 Future Scope of Work

Future research should prioritize longer-term and multi-jurisdictional collection of 911 call data to support longitudinal analyses. Integrating PulsePoint data with EMS, hospital, and trauma registry records would further improve injury severity classification and completeness. Additional work is needed to operationalize these data sources into standardized methodologies for the development of HINs and decision-support tools for practitioners. Developing guidance, visualization platforms, and automated pipelines would enable MPOs and local agencies to incorporate more equitable and comprehensive data into Vision Zero planning. Ultimately, improving data completeness is essential to ensuring that safety investments align with the true distribution of injury risk.

# Bibliography


ACS. n.d. “National Trauma Data Standard (NTDS).” Accessed December 11, 2025. https://www.facs.org/quality-programs/trauma/quality/national-trauma-data-bank/national-trauma-data-standard/.

Amoros, Emmanuelle, Jean-Louis Martin, and Bernard Laumon. 2006. “Under-Reporting of Road Crash Casualties in France.” *Accident Analysis & Prevention* 38 (4): 627–35.

“Analysis Neighborhoods | DataSF.” n.d. Accessed November 16, 2025. https://data.sfgov.org/-/Analysis-Neighborhoods/p5b7-5n3h.

Ben, Marta García, and Víctor J Yohai. 2004. “Quantile–Quantile Plot for Deviance Residuals in the Generalized Linear Model.” *Journal of Computational and Graphical Statistics* 13 (1): 36–47. https://doi.org/10.1198/1061860042949_a.

“Bucking National Trends, City of Boston Marks Progress on ‘Vision Zero.’” 2021. *StreetsblogMASS*, May 10. https://mass.streetsblog.org/2021/05/10/bucking-national-trends-city-of-boston-marks-progress-on-vision-zero/.

Calma, Emilia, and Charlotte Jackson. 2021. *Observed Disparities between 911 Calls and Crash Reports*. White Paper. DC Policy Center. https://www.dcpolicycenter.org/publications/crash-report-disparities/.

Elvik, Rune, and Anne Mysen. 1999. “Incomplete Accident Reporting: Meta-Analysis of Studies Made in 13 Countries.” *Transportation Research Record* 1665 (1): 133–40.

“Facts and Figures | Cal Poly.” n.d. Accessed December 16, 2025. https://www.calpoly.edu/about/facts-and-figures.

Flannagan, Carol A., Jonathan D. Rupp, and N. Clay Mann. 2021. *Development of a Comprehensive Approach for Serious Traffic Crash Injury Measurement and Reporting Systems*. https://trid.trb.org/View/1872874.

“Fremont Vision Zero | City of Fremont Official Website.” n.d. Accessed February 15, 2022. https://www.fremont.gov/2594/Fremont-Vision-Zero.

“Fremont Vision Zero 2020 | City of Fremont Official Website.” n.d. Accessed October 24, 2019. https://fremont.gov/2594/Fremont-Vision-Zero-2020.

“High Crash Network Streets and Intersections | Portland.Gov.” n.d. Accessed February 16, 2022. https://www.portland.gov/transportation/vision-zero/high-crash-network.

Hosseinzadeh, Aryan, Abolfazl Karimpour, Robert Kluger, and Raymond Orthober. 2022. “Data Linkage for Crash Outcome Assessment: Linking Police-Reported Crashes, Emergency Response Data, and Trauma Registry Records.” *Journal of Safety Research* 81: 21–35.

“Index of /Geo/Maps/DC2020/PUMA/St06_ca.” n.d. Accessed November 16, 2025. https://www2.census.gov/geo/maps/DC2020/PUMA/st06_ca/.

Janstrup, Kira H., Sigal Kaplan, Tove Hels, Jens Lauritsen, and Carlo G. Prato. 2016. “Understanding Traffic Crash Under-Reporting: Linking Police and Medical Records to Individual and Crash Characteristics.” *Traffic Injury Prevention* 17 (6): 580–84. https://doi.org/10.1080/15389588.2015.1128533.

“Living in San Luis Obispo | City of San Luis Obispo, CA.” n.d. Accessed December 16, 2025. https://www.slocity.org/living.

Loo, Becky PY, and K. L. Tsui. 2010. “Bicycle Crash Casualties in a Highly Motorized City.” *Accident Analysis & Prevention* 42 (6): 1902–7.

“Map of Current Police Districts | DataSF.” n.d. Accessed November 16, 2025. https://data.sfgov.org/-/Map-of-Current-Police-Districts/wkhw-cjsf.

“Map of Traffic Crashes Resulting in Injuries | DataSF.” n.d. Accessed November 18, 2025. https://data.sfgov.org/Public-Safety/Map-of-Traffic-Crashes-Resulting-in-Injuries/kn4t-hihx.

“National EMS Information System (NEMSIS) | FHWA.” n.d. Accessed December 11, 2025. https://highways.dot.gov/safety/data-analysis-tools/rsdp/rsdp-tools/national-ems-information-system-nemsis.

“Neighborhood Traffic Management | City of San Luis Obispo, CA.” n.d. Accessed December 16, 2025. https://www.slocity.org/government/department-directory/public-works/programs-and-services/transportation-planning-and-engineering/neighborhood-traffic-management.

Noland, Robert B., James A. Sinclair, Nicholas J. Klein, and Charles Brown. 2017. “How Good Is Pedestrian Fatality Data?” *Journal of Transport & Health*, Road Danger Reduction, vol. 7 (December): 3–9. https://doi.org/10.1016/j.jth.2017.04.006.

“Pedestrian Traffic Fatalities by State: 2022 Preliminary Data | Governors Highway Safety Association.” n.d. Accessed December 11, 2025. https://www.ghsa.org/resource-hub/pedestrian-fatalities-2022-data.

PulsePoint. n.d. "PulsePoint Incident Types." Accessed November 18, 2025. https://www.pulsepoint.org/incident-types.

Pulugurtha, Srinivas S., Vanjeeswaran K. Krishnakumar, and Shashi S. Nambisan. 2007. "New Methods to Identify and Rank High Pedestrian Crash Zones: An Illustration." *Accident Analysis & Prevention* 39 (4): 800–811.

Quistberg, D. Alex, Eric J. Howard, Beth E. Ebel, et al. 2015. "Multilevel Models for Evaluating the Risk of Pedestrian–Motor Vehicle Collisions at Intersections and Mid-Blocks." *Accident Analysis & Prevention* 84: 99–111.

"San Francisco, CA | Data USA." n.d. Accessed November 16, 2025. https://datausa.io/profile/geo/san-francisco-ca/.

"San Luis Obispo de Tolosa – California Missions Foundation." n.d. Accessed December 16, 2025. https://californiamissionsfoundation.org/mission-san-luis-obispo/.

Saxton, John. 2018. *A High-Injury Network for Atlanta: How Are Severe and Fatal-Injury Crashes Concentrated on Atlanta's Streets?* April. https://smartech.gatech.edu/handle/1853/59993.

"Snapshot of the SF Economy As of May 2025." n.d. Accessed November 16, 2025. https://growsf.org/news/2025-06-11-sf-economy/.

Soltani, Shamsi, Leilani Schwarcz, Devan Morris, et al. 2022. "What Is Counted Counts: An Innovative Linkage of Police, Hospital, and Spatial Data for Transportation Injury Prevention." *Journal of Safety Research* 83: 35–44.

"TIMS - Transportation Injury Mapping System." n.d. Accessed August 1, 2024. https://tims.berkeley.edu/.

"U.S. Census Bureau QuickFacts: San Luis Obispo County, California." n.d. Accessed December 16, 2025. https://www.census.gov/quickfacts/fact/table/sanluisobispocountycalifornia/PST045224.

*Vision Zero – a History*. 2019. July 10. https://actionvisionzero.org/resources/vision-zero-a-brief-history/.

*Washington Post*. n.d. "Opinion | Ignorance Is Not Bliss for the Safety of D.C. Bicyclists and Pedestrians." Accessed February 16, 2022. https://www.washingtonpost.com/opinions/local-opinions/ignorance-is-not-bliss-for-the-safety-of-dc-bicyclists-and-pedestrians/2020/02/27/c9180e74-5276-11ea-b119-4faabac6674f_story.html.

Watson, Angela, Barry Watson, and Kirsten Vallmuur. 2015. “Estimating Under-Reporting of Road Crash Injuries to Police Using Multiple Linked Data Collections.” *Accident Analysis & Prevention* 83: 18–25.

Weijermars, Wendy, Niels Bos, and Henk L. Stipdonk. 2016. “Serious Road Injuries in The Netherlands Dissected.” *Traffic Injury Prevention* 17 (1): 73–79. https://doi.org/10.1080/15389588.2015.1042577.

# About the Authors

**Anurag Pande, PhD**

Dr. Pande is a Professor of Civil Engineering at Cal Poly. His research interests include traffic simulation, data mining applications, and observational data analysis, including in the areas of traffic safety and crashes, driver behavior, transportation resilience, and emergency evacuation. As the faculty liaison for community-based learning at Cal Poly, he has worked with Cal Poly faculty and local agencies to support projects of mutual benefit. He has co-authored close to 50 manuscripts that have been either published or are forthcoming in peer-reviewed journals. He has worked on several sponsored projects, including a study on driver behavior funded by the National Science Foundation. He was the editor of the 7th edition of the Traffic Engineering Handbook (TEH), published by ITE (Institute of Transportation Engineers). Dr. Pande received his B.Tech. in Civil Engineering from the Indian Institute of Technology Bombay in Mumbai (India) and his MS and Ph.D. in Civil Engineering from the University of Central Florida (UCF).

**Anandamayee Majumdar, PhD**

Dr. Majumdar is an Assistant Professor of Statistics at San Francisco State University. Her research interests include spatial and temporal modeling and simulation, with applications to geographically referenced data, and observational data analysis. She uses and develops data analytical tools in her research and software packages in open source. Part of her research has focused on data exploration and model validation as well as spatial kriging for missing data. She was co-PI to a "highlighted project" by the National Science Foundation in 2006–2009. She has co-authored more than 40 manuscripts that have been either published or are forthcoming in peer-reviewed journals. Dr. Majumdar received her B.Stat. and M.Stat. in Statistics from the Indian Statistical Institute in Kolkata (India), her MS in Statistics from Michigan State University, and a Ph.D. in Statistics from the University of Connecticut (UConn).

**Lucas Roman Kantorowski**

Mr. Kantorowski is a Master's student in Statistics at Cal Poly. He received his Bachelor's degree in Applied Mathematics from Cal Poly San Luis Obispo in 2024.

# MTI FOUNDER

## Hon. Norman Y. Mineta

# MTI BOARD OF TRUSTEES



## Directors

**Karen Philbrick, PhD**
Executive Director

**Hilary Nixon, PhD**
Deputy Executive Director

**Asha Weinstein Agrawal, PhD**
Education Director
National Transportation Finance Center Director

**Brian Michael Jenkins**
Allied Telesis National Transportation Security Center

MINETA TRANSPORTATION INSTITUTE