

Exploring the Impact of Built Environment Factors on Robotaxi Crash Frequency in San Francisco

Shuai Wang, Emmanuel Jiang, Youngsang Jun

This study examines how macro- and micro-scale built environment characteristics influence robotaxi crash frequency in San Francisco. Using California Department of Motor Vehicles (DMV)-reported AV collision data (2019–2025), we integrate spatial, streetscape, and socio-demographic variables at the Census Block Group (CBG) level. A Random Forest model with SHAP interpretation identifies key predictors of crash risk. Results show that while density metrics (e.g., population, building) are positively associated with crash frequency, their effects are non-linear. Visual enclosure emerges as the most influential streetscape factor, indicating that perceptual constraints may challenge AV sensor performance. Land use diversity and pedestrian infrastructure are associated with reduced crash risk, whereas socio-demographic variables such as homelessness and felony rates exhibit complex spatial patterns. These findings suggest that built environment interventions could complement existing AV safety frameworks, including Vision Zero and complete streets initiatives, and support more context-sensitive, equitable planning for autonomous mobility.

1. Background

As autonomous vehicle (AV) technology continues to evolve, particularly with the rise of autonomous ride-hailing or robotaxi, transportation planners are closely watching its potential to reshape urban mobility. In recent years, advanced robotaxi services with human safety operators have gained traction worldwide in cities like Phoenix, San Francisco, Guangzhou, and Wuhan (Carlson, 2022; Dai et al., 2023; Magramo, 2024). Notably, Cruise and Waymo have begun testing and commercializing fully autonomous (safety-operator-free) robotaxis (Carlson, 2022; Dai et al., 2023). By 2035, experts project that the share of passenger miles traveled (PMT) in robotaxi and autonomous shuttles will increase from 1 percent today to 8 percent, while PMT in private cars will drop by about 15%. It will also impact transit demand, road use, and infrastructure priorities (McKinsey & Company, 2023).

2. Literature Review

While previous studies have extensively investigated the impact of built environment variables on conventional transportation safety (Zhang et al., 2024), few studies have explored how these parameters affect the crash frequency of robot cabs. This literature analysis combines previous research on macro and micro built environment elements and their impact on self-driving car and pedestrian crashes to provide a foundation for analyzing robot cab safety in San Francisco.

2.1 Macro-Level Built Environment Factors and Crash Frequency

2.1.1 Urban Density and Land Use

Existing research suggests that urban density and land-use diversity have a major impact on traffic safety. Higher population, building, and work density tend to

increase traffic exposure, resulting in increased collision risk (Dong et al., 2023). Mixed-use areas, which include residential, business, and recreational zones, are related with lower vehicle speeds and improved pedestrian safety due to heightened driver caution and walkability incentives (Wang & Vermeulen, 2021). Research on pedestrian safety backs up these findings, demonstrating that higher commercial land-use density correlates with more pedestrian crashes, whereas residential areas had lower crash risks. However, the impact of land use on AV safety is questionable. Given that AVs must strictly obey traffic laws and may encounter human driver unpredictability in mixed-use areas, it is critical to research whether robotaxis have more or fewer crashes in high-density commercial zones than conventional cars.

2.1.2 Street Design and Road Infrastructure

According to research, street network layout such as intersection density, street width, and roadway class directly affects the occurrence of crashes (Dai et al., 2024). An increase in intersection density increases crash risk due to frequent stops, turns, and pedestrian crossings. On the other hand, a well-designed grid-like roadway network with controlled intersections reduces the severity of crashes (Dong et al., 2023). Additionally, distance from transportation hubs can affect collision frequency. Pedestrian traffic and self-driving car-human interactions often increase in areas near public transportation stops, which may pose additional safety concerns for robotic cabs. Studies have shown that pedestrian-related crashes occur more frequently in high-traffic areas, meaning that robotic cabs deployed in these areas may require additional pedestrian recognition and avoidance mechanisms.

2.2 Micro-Level Built Environment Factors and Street Safety

2.2.1 Streetscape Design and Pedestrian Safety

Micro-scale urban design features such as sidewalk widths, crosswalks, and intersection geometries have an impact on both pedestrian safety and robot cab crash risk. Zhang et al. (2024) found that small sidewalks and poor pedestrian infrastructure resulted in higher pedestrian injury rates in school zones. These findings suggest that poorer pedestrian infrastructure may raise similar safety concerns at self-driving vehicle deployment sites. In addition, green infrastructure and street trees have been shown to reduce vehicle speeds while improving pedestrian safety (Dong et al., 2023). However, from the perspective of self-driving cars, too much greenery may obscure sensor visibility, making target identification difficult and increasing the likelihood of near-collisions or emergency stops in heavily vegetated locations (Chen et al., 2024).

2.2.2 Street View Imageability and Road Perception

Recent research has highlighted the importance of perceived urban aesthetics, openness and closure in shaping driver behavior. Highly figurative streets defined by salient features, visual coherence, and clear lane markings are associated with lower crash rates (Li et al., 2018). However, visually cluttered urban environments, such as locations with complex signage, numerous billboards, or an overabundance of street furniture, may confuse the audiovisual perceptual system, leading to misinterpretation and delayed responses in real-time navigation (Dai et al., 2024). In addition, lighting conditions and nighttime visibility are important variables that affect audiovisual

safety. According to research, poorly illuminated roads increase the probability of pedestrian accidents because motorists' vision is restricted. Given that self-driving car sensors use both camera-based vision and LiDAR for navigation, further research is needed to assess whether low-illumination environments disproportionately affect the perceived accuracy and crash rates of self-driving cars.

2.3 Spatial and Socioeconomic Influences on Crash Frequency

2.3.1 Spatial Heterogeneity in Built Environment and AV Safety

According to studies using spatially weighted regression models, the impact of the built environment on crash risk varies by city. Geographically weighted regression (GWR) models indicate that the impact of road network density and land use diversity on safety varies by neighborhood characteristics. This implies that spatial modeling approaches should be used to analyze macro- and micro-level built environment variables to explain local variations in crash incidence (Chen et al., 2024).

2.3.2 Socioeconomic Factors and Traffic Safety

In addition to physical infrastructure, socioeconomic factors also influence differences in automobile crash risk. Low-income neighborhoods tend to have higher pedestrian crash rates due to inadequate infrastructure improvements and limited pedestrian-friendly streetscapes (Zhang et al., 2024). Additionally, property values and median household income may be associated with roadway safety, as wealthier communities benefit from better-maintained infrastructure and traffic calming measures (Li et al., 2018). For robotaxi deployment, this means that robotic axles operating in low-income neighborhoods may face additional road risks due to poor street conditions, fewer crosswalks, and increased jaywalking. Understanding these socioeconomic disparities is critical to developing AV legislation and safety actions that promote equitable access to safe autonomous transportation.

3. Research Question

How Built Environments Shape Robotaxi Crashes in San Francisco?

4. Data and Methodology

4.1 Study Area

This study focuses on robotaxi crash data in San Francisco County, California. Its status as an early and extensive adopter of robotaxi services, combined with its diverse and complex urban fabric, makes it an ideal location for investigating the interplay between AVs and the built environment. The unit of analysis is the Census Block Group (CBG).

4.2 Data Collection

4.2.1 Dependent Variable: Robotaxi crash density

Robotaxi crash count was derived from collision reports (as PDF format) involving AVs submitted by manufacturers to the [California DMV](#), as mandated. Data covered the period from January 2019 to January 2025. Geocoded crash locations ([parsing.py](#)) were aggregated to the CBG level to calculate crash density ([df_final_Sci3.geojson](#)).

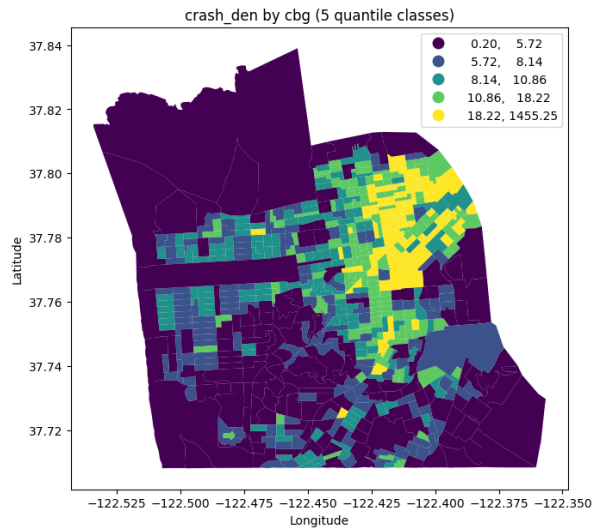


Figure 1. Crash density in San Francisco per each CBG (count/km²)

Figure 2. Report of traffic collision involving an AV (CADMV)

4.2.2 Independent Variables: Built Environment and Socio-Demographics

A comprehensive set of independent variables was compiled, categorized as follows: (Refer to Appendix A for detailed variable descriptions)

- Macro-scale built environment features (5D): Capturing broader urban form and function based on the "5D" framework.
 - Density: Population density, building density, intersection density, road network density, traffic signal density, etc.
 - Diversity: Land use mix (e.g., proportions of Commercial, Industrial, Residential land uses; Gini_Simpson index), commercial POI density.
 - Design: Tree density, open space area proportion, mean street slope, mean elevation.
 - Destination accessibility: (Partially captured through POI densities).
 - Distance to transit: Bus stop density, bus line density, metro/subway station density, parking meter density.
- Micro-scale streetscape quality features (Street View Index, SVI): Quantifying perceptual attributes of the street environment using Street View Indices derived from street-level imagery.
 - Openness/enclosure: Sky view index, enclosure index.
 - Greenness: Green view index / vegetation percentage.
 - Walkability: Spatial walkability / sidewalk percentage.
 - Other: Street furniture percentage, street obstacles percentage.
- Socio-demographic features:
 - Economic: Median household income, median property value, median

- rent, unemployment rate.
- Demographic: Population density, proportions of White, Black, Asian populations, ethnic diversity.
- Social: Homeless density, crime densities.

4.3 Analytical Framework

The analysis proceeded through the following stages:

4.3.1 Variable Filtering

To address potential multicollinearity among independent variables, Variance Inflation Factor (VIF) scores were calculated. Variables exhibiting high VIF values (e.g., > 10) were systematically removed to ensure model stability and interpretability.

4.3.2 Model Development

1. Ordinary Least Squares (OLS) Regression: An OLS model was initially fitted to serve as a baseline, examining linear relationships between the filtered independent variables and robotaxi crash density.
2. Random Forest (RF) Regression: Recognizing the potential for non-linear relationships and complex interactions, a RF model was employed as the primary analytical tool. RF, an ensemble method, is robust to outliers and capable of capturing intricate patterns while providing measures of feature importance.

4.3.3 Model Evaluation and Interpretation

1. Model Performance: Models were evaluated using standard metrics including R-squared (R^2), Mean Squared Error (MSE), and Mean Absolute Error (MAE).
2. Feature Importance: RF provides an inherent measure of feature importance based on contribution to model accuracy.
3. SHAP (SHapley Additive exPlanations): To interpret the RF model's predictions, the SHAP framework was utilized. SHAP values quantify the marginal contribution of each feature to individual predictions, enabling both global interpretation (e.g., overall feature importance, summary plots) and local interpretation (e.g., dependence plots showing how changes in a feature's value impact predictions).

5. Result

5.1 Independent Variable Filtering

Following VIF analysis, 24 independent variables were retained for modeling. The descriptive statistics and final VIF values for these variables are presented below, indicating that multicollinearity was adequately addressed (all VIF < 7.5).

Table 1: Descriptive Table for final 24 independent variables

Variables	Mean	std	VIF
<u>Socio-Demographic</u>			
Population Density	13543.798	12361.668	2.7
Whites %	43.756	21.528	7.6
Black %	4.765	8.043	1.8
Asian %	33.992	21.328	5.8
Ethnic Diversity	0.577	0.138	2.1
Median Household Income	150413.627	61399.407	1.9
Felony Density	650.637	1358.800	2.1
Homeless Density	479.902	1205.658	1.6
<u>Macro-scale Built Environment</u>			
Building Density	0.360	0.109	4.9
POI Density	4978.382	6126.947	2.8
Parking Meter Density	1723.467	2794.020	2.7
Gini Simpson Index	0.285	0.200	1.6
Intersection Density	507.100	317.792	1.7
Open Space Density	6.750	16.282	1.2
Elevation	57.309	39.255	2.1
Street Slope	5.903	3.315	2.4
Avg. Speed Limit	21.324	10.436	1.8
Transit Stop Density	120.611	113.759	1.4
<u>Micro-scale Streetscape</u>			
Visual Enclosure	0.277	0.155	6.2
Visual Walkability	0.055	0.013	2.2
Visual Motorization	0.426	0.014	2.5
Visual Vegetation	0.068	0.036	1.7
Visual Obstacle	0.004	0.004	1.9
Visual Furniture	0.004	0.001	2.1

5.2 Evaluation of Model Performance

5.2.1 OLS Model

The baseline OLS regression model achieved an R-squared of 0.718 and an adjusted R-squared of 0.706. The F-statistic was highly significant ($p < 0.001$), indicating that the selected variables collectively explain a substantial portion of the variance in robotaxi crash density.

Although the OLS model shows a high R-squared, it's potentially less suitable for this analysis. OLS assumes linear relationships, likely oversimplifying the complex, non-linear effects of the built environment on crashes which RF can better capture. Furthermore, the table indicates Covariance Type: nonrobust, meaning the model's significance tests (p-values) could be unreliable if the underlying assumption of constant error variance is violated.

OLS Regression Results			
Dep. Variable:	crash_den	R-squared:	0.718
Model:	OLS	Adj. R-squared:	0.706
Method:	Least Squares	F-statistic:	62.91
Date:	Wed, 30 Apr 2025	Prob (F-statistic):	9.84e-146
Time:	23:19:11	Log-Likelihood:	-290.83
No. Observations:	619	AIC:	631.7
Df Residuals:	594	BIC:	742.4
Df Model:	24		
Covariance Type: nonrobust			

Figure 3. OLS Regression Results

5.2.2 RF Model

The RF model demonstrated comparable explanatory power, achieving an R-squared of approximately 0.701 on standardized data, with an MSE of 0.123 and MAE of 0.268. Given its ability to handle non-linearities, subsequent interpretation focuses primarily on the RF model.

5.3 Feature Importance

The feature importance scores derived from the RF model (Figure 4) highlight the relative influence of different variables. SVI Enclosure (SVI_Enclos) emerged as the most influential predictor, followed by Population density (Population) and Building density (Building_d). Other variables with notable importance include tree density (tree_densi), felony density (Felony_den), commercial POI density (commerci_1), intersection density (intersecti), and sidewalk percentage (SVI_sidewa).

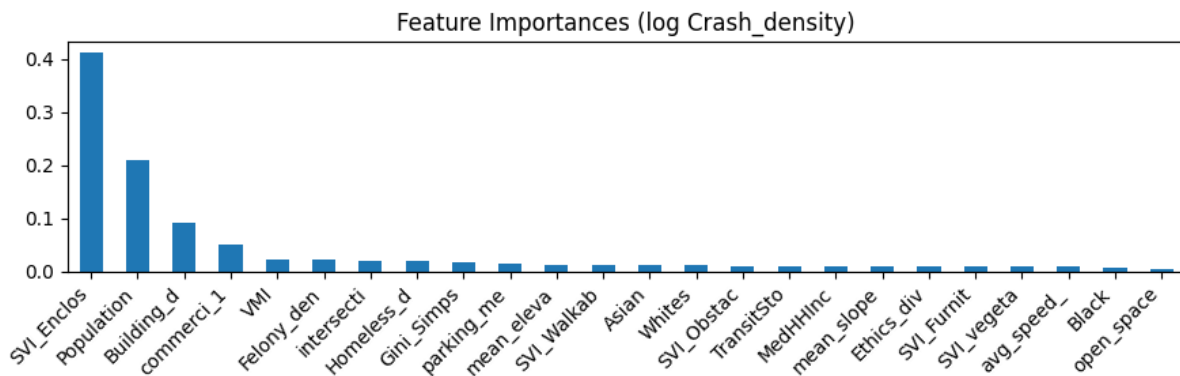


Figure 4. RF Feature Importance

5.4 SHAP Analysis: Interpreting Built Environment Impacts

SHAP analysis provided deeper insights into how each feature influences the RF model's predictions of robotaxi crash density.

5.4.1 SHAP Summary Plot

Figure 5 presents the SHAP summary plot, illustrating the global importance and impact direction of each feature on the RF model's prediction of crash density. Features are ranked vertically by importance (mean absolute SHAP value), with Visual Enclosure clearly being the most influential predictor. Each point represents an observation (CBG), where the horizontal position indicates the SHAP value (impact on prediction) and the color signifies the feature's value (red for high, blue for low). High values of Visual Enclosure (red dots) predominantly correspond to large positive SHAP values, indicating a strong positive association with increased predicted crash risk. Following Visual Enclosure, Population Density, Building Density, POI Density, and Visual Motorization are also highly ranked, generally showing that higher values (red dots) tend to increase the predicted crash risk (positive SHAP values). Conversely, features like Gini Simpson Index (Land Use Diversity) and Asian % show that higher values (red dots) tend to decrease the predicted risk (negative SHAP values), while others like Visual Obstacle and Ethnic Diversity exhibit weaker overall impacts hovering around zero.

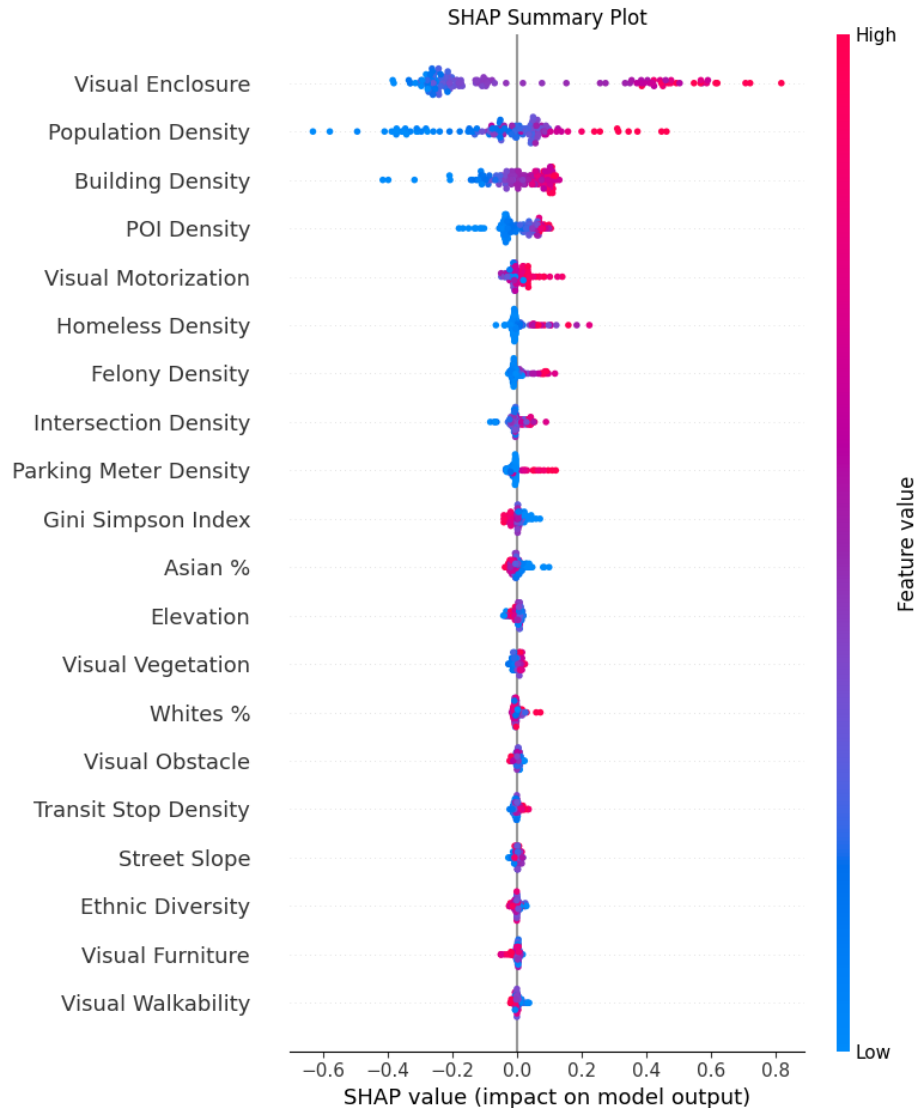


Figure 5. Results of SHAP feature importance analysis

5.4.2 Non-linear Relationships (from SHAP Dependence Plots): Examining SHAP dependence plots reveals specific patterns

1. Macro-scale Built Environment (5D) Impacts (Figure 6)

■ Density

The density-related built environment variables (X1: Building Density, X2: POI Density, X3: Parking Meter Density) all exhibit non-linear relationships with the predicted crash density, rather than simple linear associations. Specifically, Building Density (X1) and POI Density (X2) generally show a positive correlation where higher density relates to higher predicted risk, however, this effect is not constant, as the rate of risk increase slows considerably at higher density levels, indicating diminishing marginal risk effects. The relationship for Parking Meter Density (X3) is more complex, showing minimal impact at lower densities but transitioning to a significant positive association with predicted crash risk only after exceeding a certain threshold (around 4000-6000).

■ Diversity

The Gini-Simpson Index (X4), which measures built environment diversity (how mixed land uses are), shows a non-linear negative relationship with predicted crash density. As the index increases (meaning more land use mix), the SHAP value decreases, indicating that higher diversity is linked to lower predicted crash risk. This downward trend is fairly consistent across the range, suggesting a continuous safety benefit as diversity increases.

■ Design

Intersection Density (X5) shows a strong positive association, where predicted risk increases sharply at lower densities and continues to rise, albeit more slowly, at higher densities. Open Space (X6) exhibits a slightly positive, near-linear trend, suggesting marginally higher predicted risk with more open space, which might be counter-intuitive and warrants further investigation. Elevation (X7) displays a distinct non-linear pattern, with predicted risk being lowest at very low and very high elevations, peaking at moderate elevation levels (around 50-100 units). Lastly, Street Slope (X8) demonstrates a relationship that is relatively flat but shows a slight overall positive trend, suggesting that steeper slopes are associated with a marginally higher predicted crash density.

■ Accessibility/ Distance to Transit:

Average Speed Limit (X9) exhibits a relationship that is mostly flat, indicating small impact on predicted crash density within this model for speeds above approximately 25 mph. However, there is a noticeable non-linear trend where very low speed limits (below ~20 mph) are associated with negative SHAP values (lower predicted risk); this suggests that very low speed limits might indeed reduce crash probability. Transit Stop Density (X10)

displays a clear non-linear positive association: at low densities, the impact is negligible or slightly negative, but beyond a threshold of approximately 150 stops/area unit, higher density is consistently linked to an increasing predicted crash risk, likely reflecting greater pedestrian activity and complex traffic interactions near transit hubs.

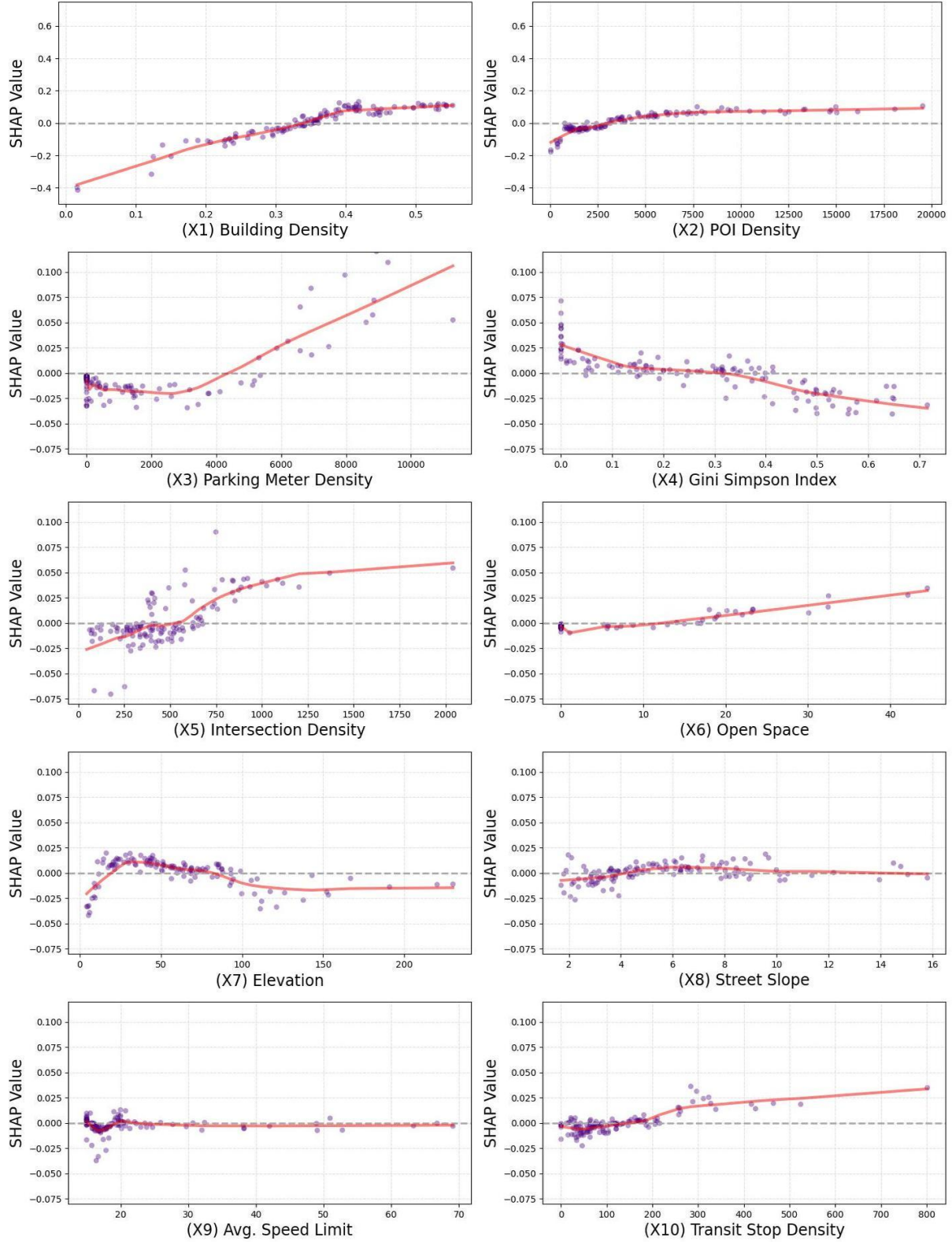


Figure 6. Non-linear relationships between crash density and macro-scale built environment variables.

2. Micro-scale Streetscape Quality (SVI) Impacts (Figure 7)

■ Enclosure/Constraint

Visual Enclosure (X11), identified as a strongly predictive variable contributing significantly to feature importance (~40%), exhibits a distinct positive sigmoidal relationship. At low enclosure levels (below ~0.2), its impact on predicted risk is minimal; however, there is a sharp and substantial increase in predicted risk as enclosure rises between approximately 0.2 and 0.4, after which the impact plateaus at a high positive level. This indicates a critical threshold beyond which increased visual enclosure drastically elevates predicted crash risk. Visual Motorization (X13) presents another complex non-linear pattern: predicted risk slightly decreases at lower motorization levels, then rises sharply within a specific intermediate range (~0.42 to 0.44), before leveling off again at higher values, suggesting a particular band of visible vehicle presence is most strongly associated with increased risk.

■ Openness/Pedestrian Environment

Visual Walkability (X12) displays a slightly negative trend, suggesting that while more visible walking space might reduce predicted risk. Both Visual Obstacle (X15) and Visual Furniture (X16) exhibit very weak, mostly flat relationships hovering close to zero impact across their observed ranges, indicating they have minimal influence on predicted crash density within this model's framework.

■ Greenness

Visual Vegetation (X14) shows a generally positive association, where predicted risk tends to increase as vegetation becomes more prominent, particularly up to a level of around 0.125, after which the effect seems to stabilize.

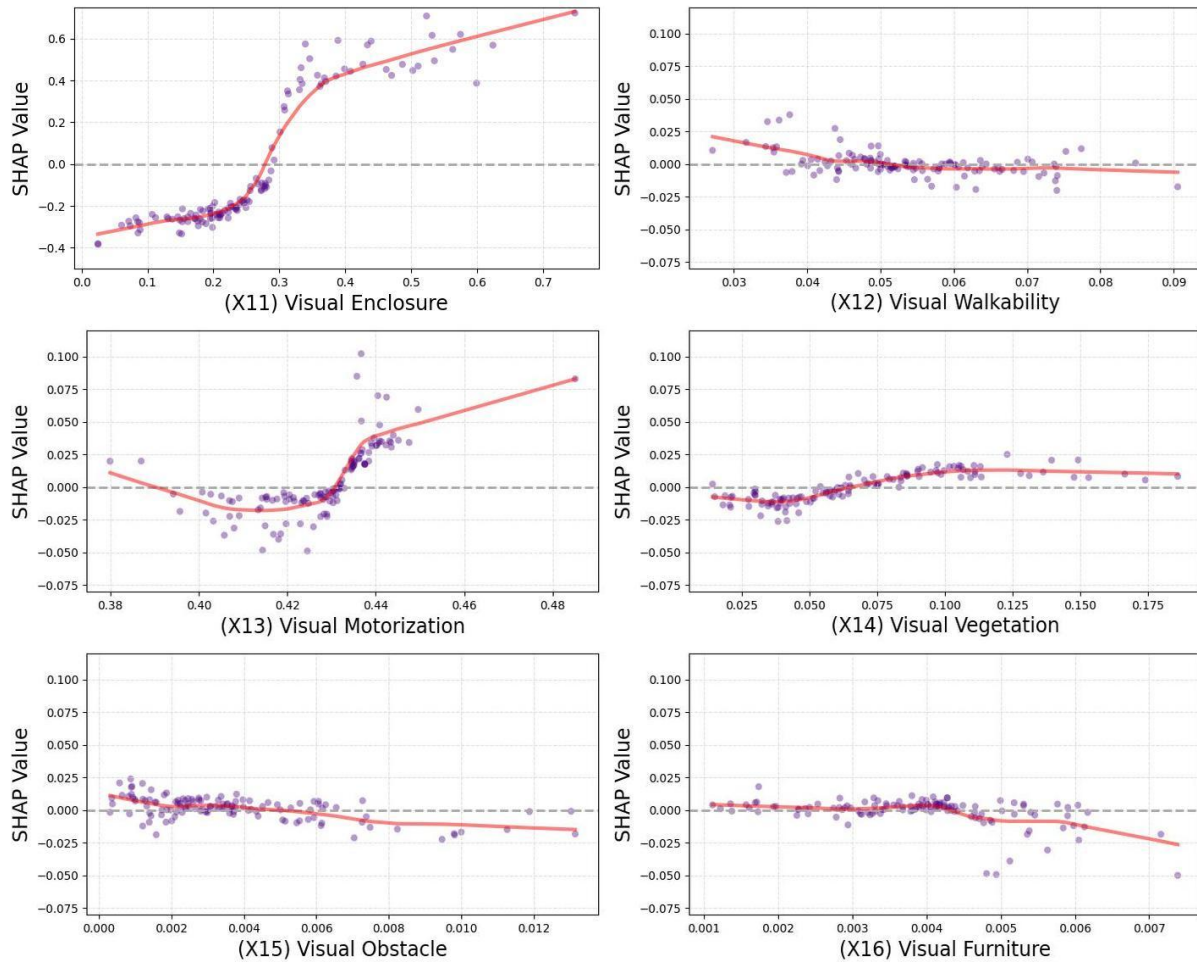


Figure 7. Non-linear relationships between crash density and streetscape variables

3. Socio-Demographic Impacts (Figure 8)

■ Demographic

Socio-demographic variables show varied non-linear impacts. Population Density (X17) has a strong positive association with predicted risk, rising sharply initially then more slowly at higher densities, suggesting diminishing marginal risk. Ethnic composition relationships are weaker and complex: Whites % (X18) is mostly flat with minor dips and rises; Black % (X19) shows minimal negative impact; Asian % (X20) has a more consistent negative association that may level off; and Ethnic Diversity (X21) appears largely unrelated to predicted risk in this model. These demographic links should be interpreted cautiously, likely reflecting correlations with unmeasured factors.

■ Economic/Social

Homeless Density (X24) shows a positive non-linear trend, where predicted risk increases as density rises from zero but the rate of increase slows considerably at higher densities (above ~2000-4000), eventually flattening out. These patterns suggest that while higher levels of crime and

homelessness density are associated with increased predicted crash risk, the marginal impact is not constant and may lessen or change direction at very high levels. Similarly, Felony Density (X23) shows a strong positive link, peaking at moderate densities before diminishing slightly. In contrast, Median Household Income (X22) exhibits a very weak, mostly flat relationship with predicted crash density. The SHAP values remain close to zero across the income spectrum, suggesting that, within this model, income level has minimal direct influence on the predicted risk after accounting for other factors.

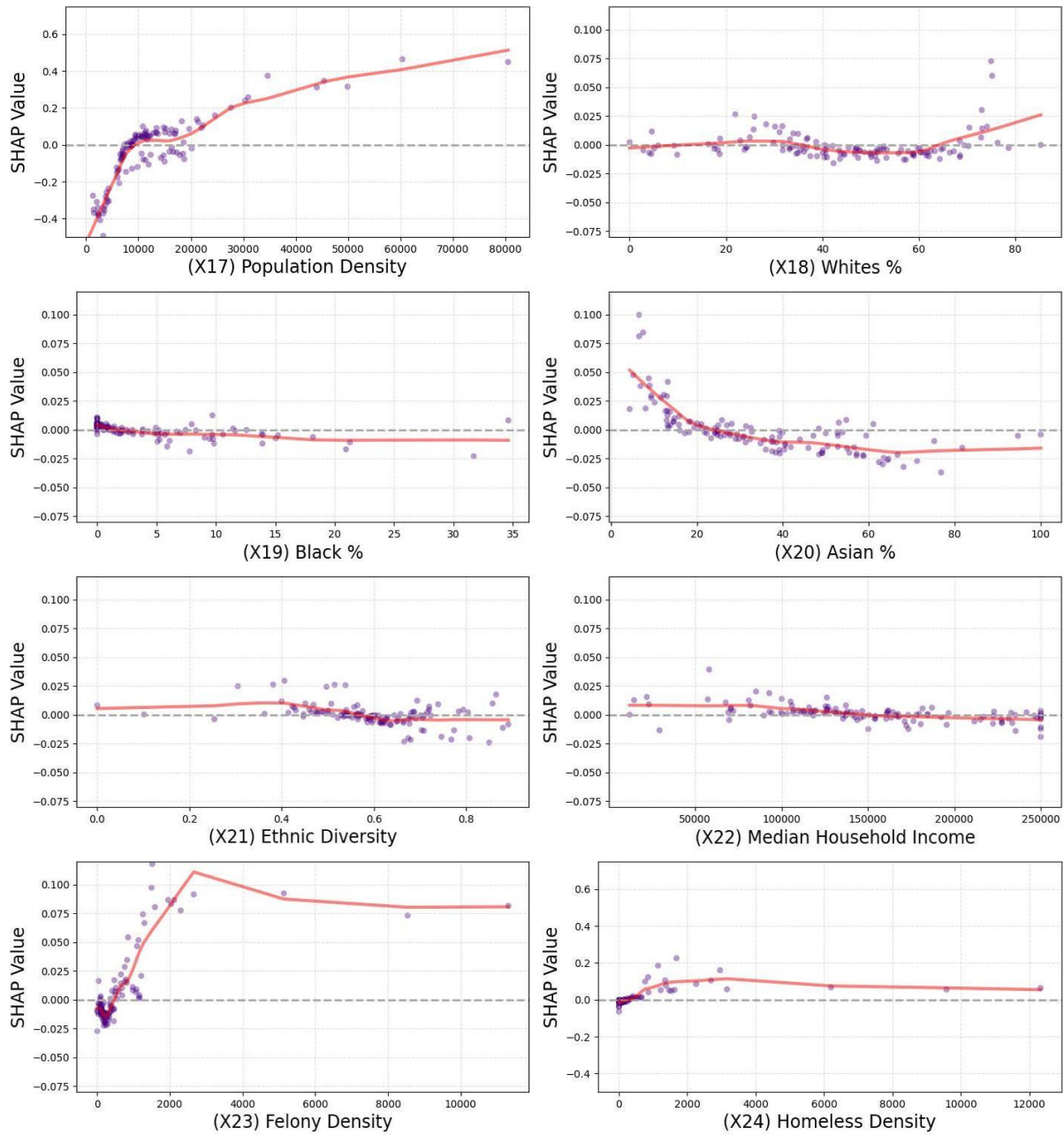


Figure 8. Non-linear relationships between crash density and socio-demographic variables.

5.5 Web Application

Using the results of this study, [the web application](#) is developed to provide information to the public sector, particularly in the planning and safety department, considering robotaxi incidents in their work. Unlike traditional PDF format incident reports, this application allows users not only to understand the spatial distribution of robotaxi crashes but also to manage and predict these incidents alongside other urban variables considered in this study. The home screen displays 600 points representing robotaxi incidents from CADMV and 676 polygons representing CBGs in San Francisco, as shown in Figure 9.

The search box in the top left enables users to find crashes at specific locations. They can search for an address, and the app allows them to zoom into the map. The incident list below the search box automatically filters to show only crashes within the current view. In the Incident List, users can edit, delete, or add a new incident by clicking any location on the map. Any newly created, edited, or deleted incident will reflect on the map and in real-time on Firebase.

Moving on to modeling, toggle switches in the layer list allow users to view the spatial distribution of each independent variable. Additionally, there are three crash density prediction scenarios based on a RF model: i) Predicted crash density if the Gini Simpson Index increases by 10%, ii) Predicted crash density if the SVI Walkability decreases by 10% (count/km²), and iii) Predicted crash density if the SVI Visual Furniture increases by 10% (count/km²). These scenarios enable planners to consider the impact of robotaxi incidents as they modify planning elements within each block group. In the top right corner, the radar chart illustrates how each block group scores on the Top-4 SHAP-contributing variables: SVI Enclosures, Population, Building Density, and Commercial POI Density. A larger radar area indicates a higher predicted crash risk for that area.

Finally, the "Download Data" button next to it allows users to filter crash data by robotaxi brand and vehicle year, which can then be downloaded in JSON format.

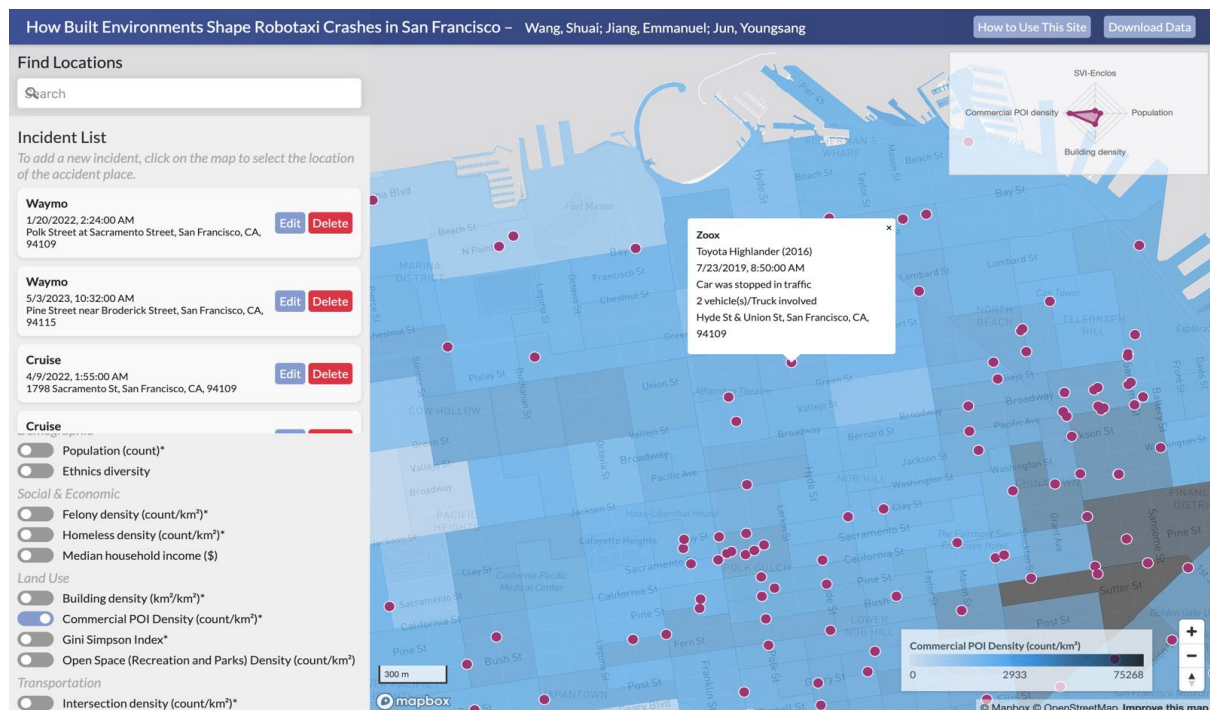


Figure 9. Web application screenshot

6. Discussion

This study investigated the influence of built environment factors on robotaxi crash frequency in San Francisco using DMV collision data and a range of environmental and socio-demographic variables. By employing RF modeling and SHAP interpretation, we identified key attributes associated with predicted crash risk.

6.1 Summary of Key Findings

6.1.1 Density Metrics Consistently Increase Predicted Risk, but Non-Linearly

Various density measures (Population, Building, Intersection, Transit Stop densities) are positively associated with higher predicted crash risk, aligning with the expectation that denser areas foster more interactions and potential conflicts. However, the SHAP analysis reveals these relationships are distinctly non-linear, often exhibiting diminishing marginal risk increases at very high densities, indicating density's impact isn't uniformly proportional across its range.

6.1.2 Visual Enclosure is the Primary Micro-Scale Risk Factor, Linked to Macro-Density

The micro-scale streetscape feature 'Visual Enclosure' emerged as the single most important predictor in the model (accounting for ~40% of feature importance). It exhibits a positive sigmoidal relationship with predicted risk: beyond a threshold (~0.2-0.4), predicted crash risk increases sharply and substantially. This high visual enclosure (often implying a high building-to-road-space ratio) is itself closely correlated with the macro-scale high-density features mentioned in the first key finding (like building and population density), commonly found in urban core areas. Therefore, high enclosure signifies not only more concentrated population and travel activity (increasing baseline exposure to potential collisions), but we also speculate that this 'street canyon' environment poses direct challenges to AV sensors (e.g., limited sightlines, signal interference), further elevating the predicted crash risk. This finding underscores the need for urban planners designing AV-friendly environments to address and mitigate excessive visual enclosure, potentially by enhancing street openness and visibility.

6.1.3 Built Environment Diversity and Walkability Show Protective Associations

Contrary to density metrics, higher built environment diversity (measured by the Gini-Simpson Index, X4) consistently correlates with lower predicted crash risk across its range, suggesting potential safety benefits inherent in mixed-use environments. Furthermore, prioritizing pedestrian infrastructure also appears beneficial for reducing predicted risk. Greater 'Visual Walkability' (X12), representing the percentage of sidewalk space, shows a negative association with predicted crash frequency, although this effect is less pronounced than that of diversity. This finding underscores the importance of robust pedestrian infrastructure – specifically providing ample and well-defined walking space like wider sidewalks – in creating street environments where predicted robotaxi crashes are less likely. The 'Visual Furniture' (X16) shows the same trend, indicating comprehensive pedestrian infrastructure contributes to overall street safety and predictability.

6.1.4 Socio-Demographic Factors Exhibit Complex and Varied Associations

Beyond Population Density (a positive predictor), other socio-demographic variables show intricate patterns. Higher Felony and Homelessness densities are linked to increased predicted risk, but non-linearly (peaking or flattening). Conversely, Median Household Income shows minimal direct impact in this model. Ethnic composition variables display complex and weaker associations requiring cautious interpretation.

6.2 Policy and Practical Implications

6.2.1 Supporting Vision Zero and Complete Streets through AV Risk Insights

Our findings suggest that built environment factors—particularly population and building density—are associated with increased predicted robotaxi crash risk, though in non-linear patterns. While these results are exploratory and context-specific, they may offer additional perspectives to help inform ongoing policy refinement such as San Francisco’s existing Vision Zero and Complete Streets initiatives. This could potentially complement existing tools used by planners when identifying high-risk areas. If integrated thoughtfully, these data-driven insights may support more targeted deployment of established measures—such as traffic calming infrastructure, 20-mph zones, or AV speed caps—in dense urban corridors.

6.2.2 Highlighting Built Form Features That May Warrant Further Attention

This model suggests that certain urban design elements—most notably visual enclosure, vegetation density, and curbside activity—may be associated with increased crash risk for robotaxis. While these findings require further validation, they highlight features that could be worth examining more closely in the context of AV deployment and street design.

For instance, visual enclosure demonstrated a non-linear relationship with crash risk in this model, which may reflect challenges for sensor-based navigation in visually constrained environments. Similarly, vegetation and curbside dynamics may influence AV sensor performance, depending on configuration and traffic conditions. These results suggest that planners and engineers might consider integrating visibility standards or clear sightline principles—already common in pedestrian safety guidelines—into future AV readiness discussions.

6.2.3 Reinforcing the Relevance of Micro-Scale Design in AV Planning

Our findings add to the growing recognition that micro-scale urban design elements, such as sidewalk presence, vegetation, and street openness, may influence AV navigation and safety performance. They align with broader planning concepts such as “AV Urbanism” (NACTO, 2024), which emphasize the need to consider built form in the transition to autonomous mobility.

In this regard, this study may contribute additional evidence in support of discussions around how streetscape features affect AV operation. While macro-level planning remains essential, incorporating fine-grained design attributes into AV

readiness frameworks—particularly in high-deployment areas—could further refine existing policy responses.

7. Limitations

This study contributes to a growing body of research examining the relationship between built environment features and AV safety. Nonetheless, several limitations must be acknowledged to contextualize the findings and inform future inquiry.

7.1 Limited AV Crash Data

The analysis relies on AV crash data submitted by manufacturers in compliance with California DMV regulations. While this dataset provides rare empirical access to real-world AV incidents, it is constrained in several ways. First, it includes only officially reported events, excluding minor collisions, near-misses, or edge-case sensor failures that did not trigger formal reporting. Dynamic and situational variables such as real-time traffic volumes, pedestrian flows, weather conditions, road condition (e.g., road pit), and vehicle-specific AV configurations (e.g., sensor placements, operational design domains) were excluded. The absence of such variables may result in unobserved heterogeneity or residual confounding, thus limiting the explanatory completeness of the model.

7.2 Spatial Aggregation and the Modifiable Areal Unit Problem (MAUP)

The use of CBGs as the spatial unit represents a compromise between geographic granularity and data availability. However, this choice introduces the Modifiable Areal Unit Problem (Openshaw, 1984), wherein observed spatial patterns may vary with the level of aggregation. Critical micro-scale characteristics—such as intersection design, pedestrian infrastructure, or visibility constraints—may be masked or diluted when averaged at the CBG level. While the integration of Street View Indices (SVIs) attempts to incorporate finer-scale perceptual features, these are still aggregated metrics and cannot fully resolve localized effects relevant to AV navigation and risk.

7.3 Residual Multicollinearity Among Key Predictors

Despite applying Variance Inflation Factor (VIF) filtering to reduce multicollinearity, some degree of conceptual and empirical overlap persists among certain predictors, particularly between population density, building density, and visual enclosure. This residual correlation complicates the interpretation of SHAP values, which assume feature independence. In dense urban environments, where these variables tend to co-occur, disentangling their individual effects on crash risk remains analytically challenging.

7.4 Absence of Spatially Explicit Modeling

While spatial heterogeneity is discussed in the results, the RF model used in this study does not explicitly incorporate spatial autocorrelation or local spatial effects. This omission limits the model's ability to account for geographically clustered risks or spatial dependence structures in crash data. Incorporating spatially explicit methods, such as Geographically Weighted Random Forest (GW-RF) or hybrid approaches combining GWR and machine learning, could enhance sensitivity to neighborhood-specific dynamics, albeit with increased methodological complexity and computational demand.

8. Conclusion

This study analyzed how built environment and socio-demographic factors influence robotaxi crash risk in San Francisco using RF model, and developed a web application which allows users to understand the spatial distribution of robotaxi crashes as well as to manage and predict these incidents alongside other urban variables.

There were key findings including the non-linear impact of density, the predictive role of visual enclosure, and the protective effects of walkability and land use diversity. Policy implications may emphasize AV-friendly street design, micro-scale planning, and improved visibility standards.

This study is quite timely, as robotaxis are expected to become more prevalent, with Waymo launching its service in D.C. within the year. This study utilized CADMV's consistent accident dataset, but it would be much more meaningful and applicable if future research could gather accident datasets from various cities across the U.S. or around the globe and conduct the same analysis.

References

- Biljecki, F., & Ito, K. (2021). Street-level urban morphology and its influence on pedestrian experience: A review of methods and applications using street view imagery. *Landscape and Urban Planning*, 210, 104069. <https://doi.org/10.1016/j.landurbplan.2021.104069>
- Carlson, A. (2022, August 11). California regulators approve commercial robotaxi service in San Francisco. *Reuters*. <https://www.reuters.com/technology/california-regulators-approve-commercial-robotaxi-service-san-francisco-2022-08-11/>
- Chen, Z., Hu, X., Yu, H., & Zhang, W. (2024). Assessing spatial heterogeneity in autonomous vehicle crash risk using geographically weighted machine learning models. *Transportation Research Part C: Emerging Technologies*, 155, 104263. <https://doi.org/10.1016/j.trc.2023.104263>
- Dai, J., Liang, Y., & Zhao, J. (2024). Streetscape complexity and self-driving vehicle safety: A visual analytics approach. *Accident Analysis & Prevention*, 192, 107217. <https://doi.org/10.1016/j.aap.2024.107217>
- Dai, J., Zhao, J., & Liang, Y. (2023). Urban form and autonomous vehicle performance: Emerging patterns from robotaxi deployments. *Journal of Planning Literature*, 38(1), 27–43. <https://doi.org/10.1177/08854122221120183>
- Dong, W., Wang, J., & Li, Y. (2023). Effects of built environment on road safety: Evidence from multi-source spatial data and machine learning models. *Transportation Research Part D: Transport and Environment*, 119, 103677. <https://doi.org/10.1016/j.trd.2023.103677>
- Li, H., Wang, Y., & Zhang, Y. (2018). Street visual design and traffic safety: Empirical analysis based on urban crash data and street view images. *Journal of Urban Design*, 23(4), 527–545. <https://doi.org/10.1080/13574809.2018.1424787>
- Magramo, M. (2024, February 15). In China, robotaxis go driverless as cities back AV tech. *South China Morning Post*. <https://www.scmp.com/tech/transport/article/3251430/driverless-robotaxis-hit-streets-guangzhou-and-wuhan-chinas-av-race-speeds>

- McKinsey & Company. (2023). *Autonomous driving's future: Making strategic choices in a changing landscape*.
<https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/autonomous-drivings-future-making-strategic-choices-in-a-changing-landscape>
- NACTO. (2024). *Blueprint for Autonomous Urbanism*. National Association of City Transportation Officials. <https://nacto.org/publication/bau/>
- Openshaw, S. (1984). *The modifiable areal unit problem*. Geo Books.
- SFMTA. (2020). *Better Market Street Project*. San Francisco Municipal Transportation Agency. <https://www.sfmta.com/projects/better-market-street-project>
- SFMTA. (2023a). *Valencia Street Safety Improvements*. San Francisco Municipal Transportation Agency. <https://www.sfmta.com/projects/valencia-street-improvements>
- SFMTA. (2023b). *Tenderloin Speed Limit Reduction*. San Francisco Municipal Transportation Agency.
<https://www.sfmta.com/projects/tenderloin-speed-limit-reduction>
- Walk San Francisco. (2021). *Dangerous intersections and daylighting*.
<https://walksf.org/our-work/victory-daylighting-dangerous-intersections/>
- Wang, J., & Vermeulen, F. (2021). Land use mix and street safety: Revisiting the relationship using fine-grained pedestrian crash data. *Journal of Transport and Land Use*, 14(1), 257–278. <https://doi.org/10.5198/jtlu.2021.1876>
- Zhang, Y., Dong, W., & Li, H. (2024). Built environment and pedestrian injury risk: A spatial machine learning analysis. *Cities*, 142, 104450.
<https://doi.org/10.1016/j.cities.2023.104450>

Appendix A. Data Collection List and Result of Exploratory Analysis

Category 1	Category 2	Variables	Description	Name in CSV	Data Source	count	mean	std
Socio demographics (SD)		Median household income	The total income of households	MedHHInc		620	150413.6274	61399.4068
		White Americans proportion		Whites		675	43.7559	21.5279
		African Americans proportion		Black		675	4.7649	8.0432
		Asian Americans proportion		Asian		675	33.9921	21.3277
		Native American proportion		Native_Ame		675	0.5649	2.0972
		Ethnic diversity	Simpson's diversity index of races	Ethics_div		675	0.5772	0.1381
		Median Rent	Median house rent per unit	MedRent		598	2368.388	789.3619
		Population density	residential population / total area	Population		676	13543.7982	12361.6683
		Property value	The property value was at the CT level	MedHVal		563	1434334.441	377868.0486
		Homeless density	Total homeless incidents divided by total area	Homeless_d		676	479.9015	1205.658
		Felony density	Total felony incidents divided by total area	Felony_de n		676	650.6372	1358.7998
		Misdemeanor density	Total misdemeanor incidents divided	Misdemea no		676	1043.1205	2711.4294

			by total area					
		Violation density	Total violation incidents divided by total area	Violation.		676	481.167	1635.7148
		Homeless density	Total homeless incidents divided by total area	Homeless_d		676	479.9015	1205.658
		Unemployment rate	Rate of unemployed people	Unemploy me		672	5.4139	6.1345
Macro-scale built environment (BE)	Density	Building density	Building area / total area	Building_d		674	0.3597	0.1091
		Intersection density	Number of intersections / total area	intersecti		676	507.0996	317.7923
		Road network density	Length (mi) of roads / total area	road_densi		676	32.1464	10.4633
		Traffic signal density	Number traffic signal / total area	traffic_si		676	55.3139	73.521
	Diversity	Commercial POI density	Number of commercial-related points of interest (POIs) / total area	commerci_1		676	4978.3824	6126.9465
		Public proportion	Proportion of land use designated as green area	Public		676	10.1925	16.8866
		Commercial land use proportion	Proportion of land use designated as commercial area	Commercial		676	4.7189	17.5151
		Industrial land use proportion	Proportion of land use designated as industrial area	Industrial		676	1.7151	8.9932
		Residential land use proportion	Proportion of land use designated as residential area	Residentia		676	64.1538	33.7559
		Mixed Use land use proportion	Proportion of land use designated as Mixed Use	Mixed.Use		676	19.2197	26.9177

		Gini_Simpson	1-HHI (Gini-Simpson)	Gini_Simp s		676	0.2853	0.1999
	Design	Tree density	Number of tree / total area	tree_densi		676	6114.1665	3371.0951
		Open space area	Open space area/ total area	open_space		676	6.7496	16.2824
		Street slope	The average slope degree within each CT	mean_slope		676	5.9028	3.3147
		Elevation	The average elevation degree within each CT	mean_eleva		676	57.3088	39.2553
	Distance to transit & Accessibility	Bus stop density	Number of bus stops / total area	Bus.Stop.D		676	103.1959	89.8999
		Bus line density	Length (mi) of bus lines / total area	Bus.Line.D		676	27.2088	36.8634
		Metro/subway station density	Number of metro/subway stops / total area	Metro.Stop		676	17.4152	45.2954
		Metro/subway line density	metro/subway lines (mi) / total area	Metro.Line		676	3.6615	13.2444
		Transit stop density	Number of bus and metro stops / total area	TransitSto		676	120.6111	113.7594
		Parking Meter Density	Number of parking meters / total area	parking_me		676	1723.4671	2794.0202
	Eye-level street environment (SE)	Sky view index	The percentage of sky in the images	SVI_sky	Google SVIs	676	0.3507	0.0561
		Enclosure index	The percentage of pixels indicating walls, fences, banister, and rails in buildings	SVI_Enclosures	Google SVIs	676	0.277	0.1552

		Green view index	Greenness indicates the presence of green elements within the streetscape, which is measured by the proportion of vegetation pixels in images	SVI_vegeta	Google SVIs	676	0.0676	0.0359
		Spatial walkability	The pixel percentage of walking space (sidewalks, paths, stairs, and stairway) in roadways	SVI_sidewa		676	0.023	0.0052
		Street furniture	The percentage of street furniture (such as awnings, benches, streetlights, traffic light, pole, lamp, escalator) in the images	SVI_Furnit	Google SVIs	676	0.0039	0.0012
		Street obstacles	The percentage of pixels indicating walls, fences, banister, and rails in buildings	SVI_Obsta	Google SVIs	676	0.0044	0.0035
		Visual motorization index		VMI	Google SVIs	676	0.4262	0.0144

Appendix B. Choropleth Maps of Variations in Each Independent Variable



