Exposure-Adjusted Bicycle Crash Risk Estimation and Safer Routing in Berlin

Eric Berger [*] **Edward Eichhorn** [*] **Liaisan Faidrakhmanova** [*] **Luise Grasl** [*] **Tobias Schnarr** [*]

Abstract

Accurately estimating the risk of bicycle crashes at street level requires consideration of both crash counts and cyclist exposure. However, exposure data from official counting stations is unavailable for most streets. This makes it difficult to identify streets that are dangerous. We solve this by using Strava's bike trip data to estimate the relative crash risk across street segments and junctions in Berlin. We identify those with a higher or lower than expected occurrence of crashes, and enable a routing algorithm to suggest lower-risk routes.

1. Introduction

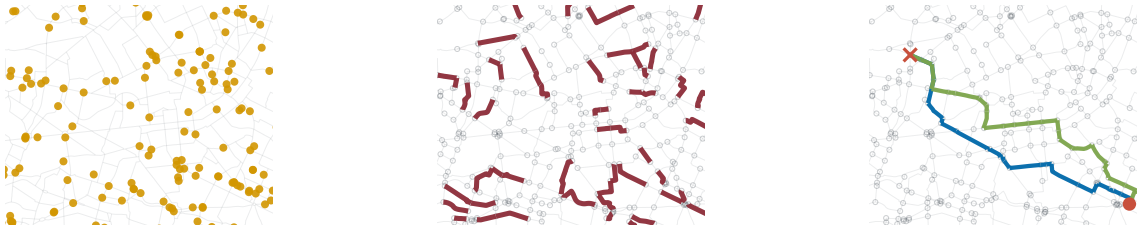
Cycling is far from a safe endeavour. 92,882 bicycle crashes were recorded in 2024, including 441 fatalities—16% of all traffic deaths that year (Destatis, 2025). Yet it is rarely clear which streets are most dangerous and thus should be avoided by cyclists or made safer. Quantifying street-level danger is non-trivial because simple crash counts confound risk with exposure. Streets with high exposure, i.e. high numbers of cyclists, tend to accumulate more crashes even when per-cyclist risk is low (Lücken, 2018). Crashes must be normalised by cyclist counts; otherwise, dangerous streets can remain hidden in dense urban networks (Uijtdewilligen et al., 2024). Unfortunately, street-level cyclist counts are rare. Berlin, for example, provides hourly counts via official counting stations, but their limited coverage (20 stations for thousands of streets) makes them impractical for city-wide risk estimation (Senate Department for Urban Mobility, Transport, Climate Action and the Environment). We address this problem by using bike trip counts from the fitness-tracking app Strava. These have been used to predict official bike counts (Dadashova et al., 2020). We show that they can serve as a proxy for cyclist exposure and estimate for all segments and junctions in Berlin's official cycling network relative risks (the ratio of observed to expected crashes). Because Strava coverage can be sparse, we use Empirical Bayes smoothing for estimation (Clayton & Kaldor, 1987). This stabilises estimates for low-exposure segments and junctions, and quantifies uncertainty. We also introduce a routing algorithm that finds substantially lower-risk routes under a route-length constraint.

[*]Equal contribution. Correspondence to: Tobias Schnarr <tobias-marco.schnarr@student.uni-tuebingen.de>.

Project report for the "Data Literacy" course at the University of Tübingen, Winter 2025/26 (Module ML4201). Style template based on the ICML style files 2025.

2. Data

Multiple datasets were used for risk estimation. Crash counts were taken from the *German Accident Atlas* (Destatis, 2025), which provides geodata of police-reported crashes where people were injured. We filtered the data to bicycle-related crashes within the city limits of Berlin. Cyclist exposure was approximated using the dataset by Kaiser et al. (2025b), which reports daily street-segment-level counts of bicycle trips recorded via the Strava app in Berlin from 2019 to 2023. Strava users are not representative of the general cycling population (they skew younger, male, and sport-oriented; Kaiser et al., 2025b). Therefore, we assess potential bias by comparing segment-level count shares in 2023 with official bicycle counter data from the city of Berlin (Senate Department for Urban Mobility, Transport, Climate Action and the Environment, 2024) for the subset of segments where both Strava and official counts are available (Figure 2). Count shares correlate strongly ($r = .61$) and are preserved in the Strava data. Segments on main streets where one can ride fast (e.g., Karl-Marx-Allee) are overrepresented in the Strava data, since those are more often tracked. Residential streets (e.g., Kollwitzstraße) are underrepresented, since slower, everyday cycling is less often tracked. All datasets were combined into one dataframe and matched to the same street network. The network is represented as segments with associated monthly exposure counts. We map crashes to the network using nearest-segment assignment. Junctions are defined as nodes where at least three segments meet and crashes within a fixed radius are assigned to the nearest junction. Their exposure is derived from segment exposure (see Section 3). At monthly resolution, events are sparse: in a typical month, fewer than 4% of segments and 3% of junctions record at least one crash. We drop segments with zero recorded trips over at least one year and pool counts over the full period 2019–2023 for risk estimation. The final dataset comprises 4,335 segments, 2,862 junctions, and 15,396 recorded bicycle crashes.



(a) Data (b) Risk estimation (c) Safety routing

Figure 1. **Safety-aware routing pipeline for the Berlin network.** Panels (a–c) are zoomed in for readability; see Section 3 for definitions and notation. (a) Police-recorded bicycle crashes in June 2021 (points) and street segments with measured cyclist exposure (lines). (b) Pooled segment-level relative risk estimated from all available data; high-risk segments in red correspond to values above the 90th percentile of relative risk; circles mark junctions (degree ≥ 3). (c) Shortest path (blue) versus a safer alternative (green) selected to reduce cumulative relative risk under a route-length constraint. Filled circle and cross denote origin and destination, respectively.

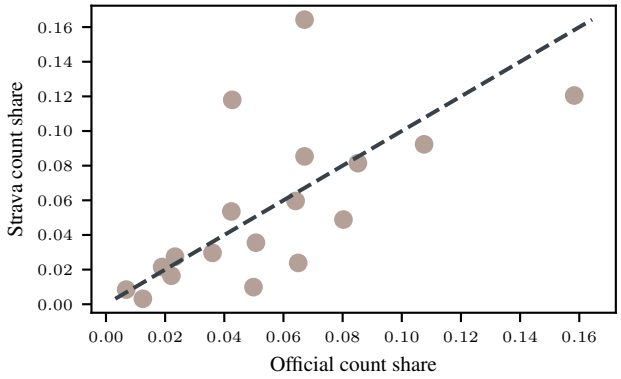


Figure 2. Consistency check between official bicycle counts and Strava bike trips at the street-segment level (2023). Points show segment-wise shares of total annual counts.

3. Methods

Crash, exposure, and risk measures. For each month t, let $C_{s,t}$ and $E_{s,t}$ denote the number of police-recorded bicycle crashes and measured cyclist exposure on street segment s. Junction crashes $C_{v,t}$ are defined as crashes within a fixed radius of junction v. Because a traversal typically contributes exposure to two incident segments, we approximate junction exposure by the half-sum of incident segment exposures,

$$E_{v,t} = \tfrac{1}{2} \sum_{s \in \mathcal{I}(v)} E_{s,t},$$

a common approach when turning movements are unavailable (Hakkert & Braimaister, 2002; Wang et al., 2020). For notational convenience, both street segments and junctions are indexed by a generic entity index i, with $A_{i,t}$ and $E_{i,t}$ denoting the corresponding crash and exposure quantities.

Under a no-special-risk baseline, crash incidence is assumed proportional to exposure, yielding the expected number of crashes and total sum of crashes for a given street-segment i aggregated over time.

$$\widehat{C}_i = \frac{\sum_{i,t} C_{i,t}}{\sum_{i,t} E_{i,t}} \sum_t E_{i,t}, \quad C_i = \sum_t C_{i,t}.$$

The raw relative risk r_i^{raw} is then given by $C_i / \widehat{C}_i$.

Empirical Bayes smoothing. Because many segments have low exposure and thus very small expected counts, the raw relative risk is highly variable. That is why we use Empirical Bayes smoothing to improve the risk estimates. This methods shrinks low-exposure estimates toward a baseline, while high-exposure estimates change little. Concretely, we assume a true relative risk r_i^{true} such that the observed count C_i follows a Poisson model with

$$C_i \mid r_i^{\text{true}} \sim \text{Poisson}(\widehat{C}_i \, r_i^{\text{true}}).$$

The Poisson distribution is natural for nonnegative event counts over a fixed time period under a baseline rate, and it yields $\mathbb{E}[C_i] = \widehat{C}_i$ when $r_i^{\text{true}} = 1$. To allow heterogeneity in relative risk beyond this baseline, we place a Gamma prior on r_i^{true} in the shape–rate parameterization,

$$r_i^{\text{true}} \sim \text{Gamma}(\alpha, \alpha),$$

which enforces $\mathbb{E}[r_i^{\text{true}}] = 1$ and has variance $\text{Var}(r_i^{\text{true}}) = 1/\alpha$ controlling the amount of shrinkage. The Gamma prior is also conjugate to the Poisson likelihood, giving a closed-form posterior

$$r_i^{\text{true}} \mid C_i, \widehat{C}_i \sim \text{Gamma}(C_i + \alpha, \ \widehat{C}_i + \alpha),$$

so posterior inference is simple and numerically stable. We estimate α from the data using method of moments (Morris, 1983), as

$$\widehat{\alpha} = \frac{\sum_i \widehat{C}_i^2}{\sum_i (C_i - \widehat{C}_i)^2 - \sum_i \widehat{C}_i}.$$

and use the posterior mean

$$\widehat{r}_i = \mathbb{E}[r_i^{\text{true}} \mid C_i, \widehat{C}_i] = \frac{C_i + \alpha}{\widehat{C}_i + \alpha}$$

as the smoothed relative risk. For small $\widehat{C}_i$, r_i is pulled toward 1, while for large $\widehat{C}_i$ it approaches the raw ratio $C_i/\widehat{C}_i$. Uncertainty is summarized by $(1 - \delta = 0.95)$ equal-tailed credible intervals from quantiles of the Gamma posterior. Since our goal is to identify segments or junctions whose risk deviates from the baseline, we first set the relative risk to 1 whenever its credible interval includes 1. For all remaining cases, we use a conservative deviation estimate: the upper credible limit if the interval lies entirely below 1, and the lower credible limit if it lies entirely above 1. We use these adjusted risks in all subsequent analyses.

Risk-weighted routing graph. Relative risk estimates are dimensionless and conditional on exposure. To obtain additive routing weights, we rescale relative risk by the pooled baseline crash rate,

$$\bar{\lambda} = \frac{C_\cdot}{E_\cdot}, \qquad C_\cdot = \sum_i C_i, \;\; E_\cdot = \sum_i E_i,$$

yielding the routing weight $w_i = \bar{\lambda}\,\widehat{r}_i$. We construct an undirected graph $G = (V, E)$ from the street network, where nodes correspond to segment endpoints and edges to street segments of length ℓ_e. Each edge e corresponds to a segment s and inherits its weight, $w_e = w_s$. Junction identifiers and weights are mapped to nodes via spatial snapping in a projected coordinate system, producing a single risk-annotated network.

Safety-aware routing. We compare shortest-distance routes with alternatives that reduce estimated crash risk under a bounded detour. The length of a route P is

$$L(P) = \sum_{e \in P} \ell_e.$$

To incorporate segment- and junction-level risk, the risk contribution of edge $e = (u, v)$ is defined as

$$\rho_e = w_e + \frac{w_u + w_v}{2},$$

where w_u and w_v denote junction routing weights (zero for non-junction nodes), yielding an additive surrogate for cumulative route risk.

For an origin–destination pair, the baseline route P_{dist} minimizes $L(P)$. The safety-aware route is obtained by solving

$$P_{\text{safe}} = \arg\min_P \; R(P) = \arg\min_P \sum_{e \in P} \rho_e$$
$$\text{s.t. } L(P) \le (1 + \varepsilon)\, L(P_{\text{dist}}),$$

where ε is the allowable relative detour (Ehrgott, 2005). We approximate this constraint using a weighted-sum sweep: for $\lambda \in \Lambda$,

$$P(\lambda) = \arg\min_P \left(\sum_{e \in P} \rho_e + \lambda \sum_{e \in P} \ell_e \right),$$

and select the feasible route minimizing $R(P)$. Shortest paths are computed using Dijkstra's algorithm (Dijkstra, 1959).

Evaluation metrics. For each origin–destination pair, we report the relative length increase and relative risk reduction, as

$$\Delta_L = \frac{L(P_{\text{safe}}) - L(P_{\text{dist}})}{L(P_{\text{dist}})}, \quad \Delta_R = \frac{R(P_{\text{dist}}) - R(P_{\text{safe}})}{R(P_{\text{dist}})}.$$

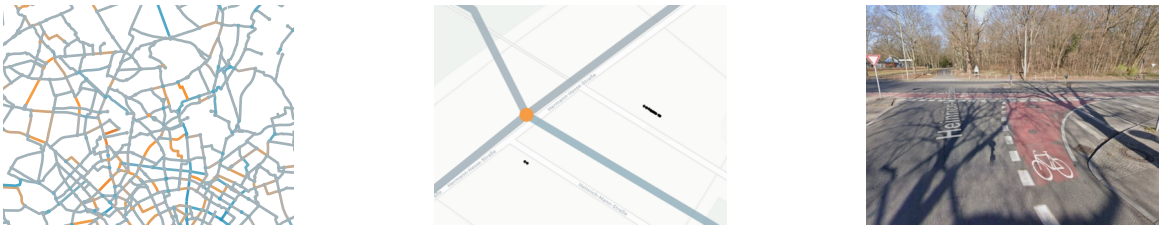
We additionally report the expected number of avoided crashes, as $\Delta_C = R(P_{\text{dist}}) - R(P_{\text{safe}})$. Pairs with $R(P_{\text{dist}}) = 0$ are excluded from Δ_R.

4. Related Work

Previous studies on estimating street-level crash risk differ mainly in the proxy used for cyclist exposure and how exposure is modelled. For instance, Wage et al. (2022) extrapolated bicycle volumes from motorised traffic data, but this does not accurately reflect actual cycling patterns. As an alternative to Strava-based exposure estimation, other studies have used crowdsourced GPS traces from less widely used tracking apps (e.g. BikeCitizens; Medeiros et al., 2021) or city-wide counting events (Uijtdewilligen et al., 2024). While these sources can approximate long-term station counts, they tend to be less stable as they rely on small user bases or special occasions. Another approach predicts bicycle counts directly with models such as XGBoost and graph neural networks (Kaiser et al., 2025a;b). While these methods can achieve strong predictive performance, they typically require rich, location-specific inputs and often additional calibration data, including manual counts on some streets. This is costly and rarely feasible at scale. To address crash sparsity and overdispersion, prior work has also used Poisson–Gamma count models (Lücken, 2018). However, this was only for city-wide crash estimation and is not suited to estimating risk at segment level.

5. Results

We computed the relative risk values ($\widehat{r}_i$) for all street segments and junctions. Due to high variance in the crash data, the shrinkage parameter is small ($\widehat{\alpha} = 0.129$), resulting in limited regularization and a wide spread of $\widehat{r}_i$ estimates. These are shown in Figure 3(a). Most elements exhibit low



(a) Risk heatmap (b) High-risk junction (c) Street-level view

Figure 3. **Risk heatmap and detailed inspection of junction 2482.** Colors in panels (a)–(b) indicate $\log_{10}$-scaled risk values, ranging from low (-2) to high (2). (a) Section bike network with all computed road segments and junctions displayed. (b) Closer view of junction 2482 and the crashes (black dots) assigned to it (c) Street-level view of junction 2482 (Google, 2025), providing visual context for the observed risk.

risk: 64.4% of segments and 69.1% of junctions lie within the confidence bounds of the baseline ($\widehat{r}_i = 1$). Values of $\widehat{r}_i < 1$ occur for 17.6% of segments and 25.8% of junctions, while 17.9% of segments and 4.9% of junctions show elevated risk ($\widehat{r}_i > 1$). Risk values range from 0.03–50.79 for segments and 0.03–6.43 for junctions. To verify that the method identifies high-risk locations, one such site was examined in detail. At junction 2482 ($\widehat{r}_i = 6.43$), 22 crashes, most of them clustered along the right turning bike lane, occurred despite moderate traffic. All involved at least one additional vehicle—20 of them were cars. This seems plausible, since—as shown in Figure 3(c)—car lanes intersect the bicycle lane at this junction.

Furthermore, we evaluated the routing algorithm using 1,000 random origin–destination pairs, comparing the shortest-path baseline against safest alternatives with a detour constraint (Natera Orozco et al., 2020).

Table 1. Evaluation of the safety-aware routing algorithm with varying relative detour budgets (ε). Values are aggregated over all origin–destination pairs and reported as medians with interquartile ranges. Δ_L denotes the relative path length increase, Δ_R the relative risk reduction, and Δ_C the expected number of avoided crashes per 100,000 trips, reported as rounded integer counts.

ε	Δ_L (IQR)	Δ_R (IQR)	Δ_C (IQR)	$\Delta_R > 0$
0.05	0.007 (0.022)	0.101 (0.210)	39 (123)	0.767
0.10	0.015 (0.037)	0.147 (0.240)	61 (132)	0.841
0.15	0.024 (0.063)	0.169 (0.239)	71 (141)	0.873
0.20	0.041 (0.109)	0.192 (0.238)	83 (150)	0.901
0.30	0.091 (0.156)	0.237 (0.228)	104 (157)	0.928
0.40	0.137 (0.188)	0.262 (0.230)	112 (170)	0.944
0.50	0.165 (0.200)	0.281 (0.222)	121 (178)	0.948

Table 1 illustrates a clear trade-off between path length and safety, where increasing the detour budget ε yields consistent safety improvements. At the smallest budget of $\varepsilon = 0.05$, the algorithm achieves a median relative crash reduction (Δ_R) of 10.1%, avoiding 39 expected crashes per 100,000 trips (Δ_C), while incurring a median distance increase (Δ_L) of only 0.7%. As the budget relaxes to $\varepsilon = 0.50$, the median safety gains scale to $\Delta_R = 28.1\%$ (121 expected avoided crashes), and the actual median path increase reaches 16.5%. While the interquartile ranges (IQR) for safety gains (Δ_R, Δ_C) remain relatively stable yet high, reflecting a large and persistent variability across routes, the IQR for path length (Δ_L) expands significantly from 0.022 to 0.200, indicating that the distance cost required to achieve these improvements is highly sensitive to the specific layout of the route. Furthermore, the likelihood of finding a safer route ($\Delta_R > 0$) remains high across all budgets, increasing from 76.7% to 94.8% as the search space expands.

6. Discussion and Conclusion

We developed a method to estimate bicycle crash risk at the street-segment level using police-recorded crash counts and exposure data derived from Strava. These risk estimates were used to identify high-risk locations, where clear hazards could be observed. Based on these estimates, we implemented a safety-aware routing algorithm that balances crash risk reduction against constraints on route length. For modest detours of 5–15%, the routing approach achieves substantial reductions in expected crash risk. The median risk decrease ranges from 10.1% to 16.9%, corresponding to an expected avoidance of 39–71 crashes compared to the shortest-path route. However, the reduction in route-level risk strongly depends on the chosen origin and destination. Safe alternative routes are available for the majority of simulated trips (76.7%–87.3%). Several limitations must be acknowledged. Official bicycle crash counts are underreported and only include crashes involving injuries. In addition, the exposure data used in this study does not represent the full cycling population due to the nature of Strava data. Finally, when computing junction exposure, we assume uniform flow and do not account for turning movements.

Code and supplementary materials are available at https://github.com/ytobiaz/data_literacy.

Contribution Statement

Eric Berger worked on visualizations and data correctness. Edward Eichhorn worked on Empirical Bayes Smoothing and investigating the risk estimates. Liaisan Faidrakhmanova worked on aggregating raw data and merging. Luise Grasl worked on preprocessing and visualizations. Tobias Schnarr worked on merging and routing. All authors jointly wrote the text of the report.

References

Clayton, D. and Kaldor, J. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43(3):671, September 1987. ISSN 0006-341X. doi: 10.2307/2532003. URL http://dx.doi.org/10.2307/2532003.

Dadashova, B., Griffin, G. P., Das, S., Turner, S., and Sherman, B. Estimation of average annual daily bicycle counts using crowdsourced strava data. *Transportation Research Record: Journal of the Transportation Research Board*, 2674(11):390–402, September 2020. ISSN 2169-4052. doi: 10.1177/0361198120946016. URL http://dx.doi.org/10.1177/0361198120946016.

Destatis. German accident atlas, 2025. URL https://unfallatlas.statistikportal.de/. Retrieved January 14 2026.

Dijkstra, E. W. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959. ISSN 0945-3245. doi: 10.1007/bf01386390. URL http://dx.doi.org/10.1007/BF01386390.

Ehrgott, M. *Multicriteria Optimization*, volume 491 of *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin, Heidelberg, 2005. ISBN 978-3-540-21398-7. URL https://doi.org/10.1007/3-540-27659-9.

Google. Google Street View: Junction Heinrich-Mann-Straße/Hermann-Hesse-Straße Berlin, 2025. URL https://maps.app.goo.gl/pxxqfSwW8Rbtu6AZ8.

Hakkert, A. S. and Braimaister, L. The uses of exposure and risk in road safety studies. Technical Report R-2002-12, SWOV Institute for Road Safety Research, Leidschendam, The Netherlands, 2002. URL http://www.swov.nl/rapport/R-2002-12.pdf.

Kaiser, S. K., Klein, N., and Kaack, L. H. From counting stations to city-wide estimates: data-driven bicycle volume extrapolation. *Environmental Data Science*, 4, 2025a. ISSN 2634-4602. doi: 10.1017/eds.2025.5. URL http://dx.doi.org/10.1017/eds.2025.5.

Kaiser, S. K., Rodrigues, F., Azevedo, C. L., and Kaack, L. H. Spatio-temporal graph neural network for urban spaces: Interpolating citywide traffic volume, 2025b. URL https://arxiv.org/abs/2505.06292.

Lücken, L. On the variation of the crash risk with the total number of bicyclists. *European Transport Research Review*, 10(2):33, 2018. doi: 10.1186/s12544-018-0305-9. URL https://doi.org/10.1186/s12544-018-0305-9.

Medeiros, R. M., Bojic, I., and Jammot-Paillet, Q. Spatiotemporal variation in bicycle road crashes and traffic volume in berlin: Implications for future research, planning, and network design. *Future Transportation*, 1(3):686–706, 2021. ISSN 2673-7590. doi: 10.3390/futuretransp1030037. URL https://www.mdpi.com/2673-7590/1/3/37.

Morris, C. N. Parametric empirical bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):47–55, 1983. ISSN 1537-274X. doi: 10.1080/01621459.1983.10477920. URL http://dx.doi.org/10.1080/01621459.1983.10477920.

Natera Orozco, L. G., Battiston, F., Iñiguez, G., and Szell, M. Data-driven strategies for optimal bicycle network growth. *Royal Society Open Science*, 7(12):201130, 2020. ISSN 2054-5703. doi: 10.1098/rsos.201130. URL http://dx.doi.org/10.1098/rsos.201130.

Senate Department for Urban Mobility, Transport, Climate Action and the Environment. Zählstellen und fahrradbarometer: Fahrradverkehr in zahlen. URL https://www.berlin.de/sen/uvk/mobilitaet-und-verkehr/verkehrsplanung/radverkehr/weitere-radinfrastruktur/zaehlstellen-und-fahrradbarometer/.

Senate Department for Urban Mobility, Transport, Climate Action and the Environment. Radverkehrszählstellen – jahresbericht 2023, 2024. URL https://www.berlin.de/sen/uvk/_assets/verkehr/verkehrsplanung/radverkehr/weitere-radinfrastruktur/zaehlstellen-und-fahrradbarometer/bericht_radverkehr_2023.pdf?ts=1752674590. Stand: 31.05.2024 (Berlin, Mai 2024). Accessed: 2026-02-01.

Uijtdewilligen, T., Ulak, M. B., Wijlhuizen, G. J., Bijleveld, F., Geurs, K. T., and Dijkstra, A. Examining the crash risk factors associated with cycling by considering spatial and temporal disaggregation of exposure: Findings from four dutch cities. *Journal of*

Transportation Safety & Security, 16(9):945–971, 2024. doi: 10.1080/19439962.2023.2273547. URL https://doi.org/10.1080/19439962.2023.2273547.

Wage, O., Bienzeisler, L., and Sester, M. Risk analysis of cycling accidents using a traffic demand model. *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, XLIII-B4-2022:427–434, 2022. doi: 10.5194/isprs-archives-XLIII-B4-2022-427-2022. URL https://isprs-archives.copernicus.org/articles/XLIII-B4-2022/427/2022/.

Wang, K., Zhao, S., and Jackson, E. Investigating exposure measures and functional forms in urban and suburban intersection safety performance functions using generalized negative binomial - p model. *Accident Analysis & Prevention*, 148:105838, 2020. ISSN 0001-4575. doi: https://doi.org/10.1016/j.aap.2020.105838. URL https://www.sciencedirect.com/science/article/pii/S0001457520316584.