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

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

Abstract

We investigate bicycle crash risk on Berlin's urban street network, addressing a key limitation of many safety analyses: raw crash counts conflate danger with demand and fail to distinguish intrinsically risky locations from high-use roads. We combine police-reported crashes with a city-wide dataset of measured bicycle volumes to compute exposure-adjusted risk at the street-segment and junction levels. Aggregating risk to arbitrary routes enables comparisons that trade off safety against convenience. The result is a reproducible framework for exposure-controlled bicycle safety analysis and routing.

/edit, when we nal results

1. Introduction

Cycling safety analyses often rely on raw crash counts, which conflate danger with demand: streets that attract many cyclists tend to accumulate more incidents even when per-rider risk is low (Lücken, 2018). This obscures intrinsically risky locations and limits both targeted interventions and everyday route choice, especially in dense urban networks such as Berlin (Uijtdewilligen et al., 2024). We address this by estimating exposure-adjusted crash risk on the street network via a relative-risk formulation that separates cyclist demand from intrinsic danger and remains stable under sparse or unevenly distributed observations (Section 3). Using police-recorded crashes from the German Unfallatlas (Berlin subset) (German Federal Statistical Office, 2025) together with a city-wide dataset of measured bicycle volumes at the street-segment level (Kaiser, 2025), we compute exposure-adjusted relative risk for individual street segments and, motivated by the concentration of crashes at intersections, derive junction-level risk by aggregating volumes from adjoining segments. The resulting risk estimates are transformed into expected crash costs and propagated to route-level scores, enabling the evaluation of arbitrary routes and comparisons between alternatives that reduce estimated risk while maintaining comparable convenience in terms of distance (Section 5), see Figure 1.

Taken together, this yields a reproducible pipeline for estimating exposure-adjusted crash risk at street and junction levels from measured cyclist volumes, and a route-scoring approach that integrates network-level risk into safety-aware routing under a bounded detour constraint.

2. Data

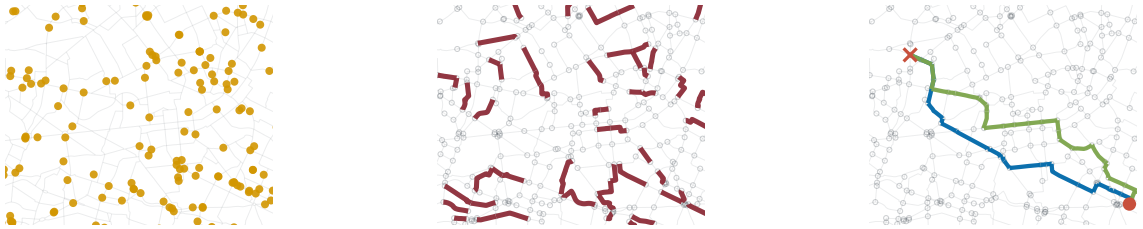
We combine police-recorded bicycle crashes with measured cyclist exposure for the city of Berlin. Crash data are drawn from the Berlin subset of the German *Unfallatlas* (German Federal Statistical Office, 2025) and filtered for bicycle-related incidents. Exposure comes from a city-wide dataset of measured bicycle volumes aggregated at the street-segment level (Kaiser et al., 2025b). We further evaluate how well cyclist volumes derived from Strava align with measurements from official bicycle counting stations (see Figure 2). The street network is represented as polyline segments with associated monthly cyclist counts. The resulting dataset spans 2019–2023 and covers 4,335 street segments and 2,862 junctions, with 15,396 recorded bicycle crashes. At monthly resolution the data are sparse: in a typical month fewer than 5% of segments and about 3% of junctions record at least one crash, and some periods include segments with zero measured exposure.

verify the percentages

To enable network-scale analysis, all layers are harmonized to a common street-network topology and a projected coordinate reference system. Crash locations are matched to segments using nearest-segment assignment. To capture the concentration of crashes at intersections, we identify junctions as nodes where at least three segments meet; crashes within a fixed radius are assigned to the nearest junction, and junction exposure is computed from the exposures of incident segments. Matched crashes and exposures are aggregated to monthly resolution, and segments and months with zero exposure are dropped. To obtain stable risk estimates under sparse observations, monthly aggregates are pooled to yearly totals and an Empirical Bayes approach is applied as described in Section 3. The resulting yearly segment- and junction-level risk estimates serve as inputs to all routing analyses.

*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.



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

Figure 1. **Safety-aware routing pipeline for the Berlin cycling network.** Panels (a–c) are zoomed in for readability; see Section 3 for formal definitions and notation. (a): Police-recorded bicycle crashes in June 2021 (points) and street segments with measured cyclist exposure (lines, used as the base network in all panels). (b): Pooled segment-level relative crash risk estimated from all available data; high-risk segments in red correspond to the top 90th percentile of relative risk. Circles mark junctions (degree ≥ 3), for which we also estimate junction-level relative risk from exposure aggregated over incident segments. (c): Shortest path (blue) versus a safer alternative (green) selected to reduce cumulative relative route risk under a distance-detour constraint. Filled circle and cross denote origin and destination, respectively; circles show junctions for reference.

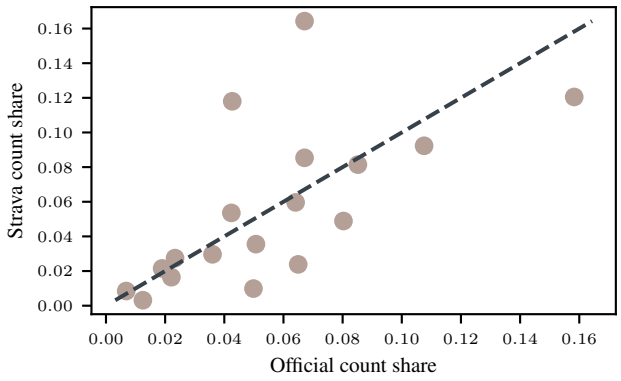


Figure 2. Alignment between official bicycle counts and Strava-derived cyclist volumes at the street-segment level in 2023. Points show segment-wise shares of total annual counts, with the dashed line indicating perfect agreement between the two sources.

3. Methods

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

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

which avoids double-counting. Aggregating exposure over incident links is common when turning movements are unavailable (Hakkert et al., 2002; Wang et al., 2020).

For notational convenience, we index both street segments and junctions by a generic entity index i, with $A_{i,t}$ and $E_{i,t}$ denoting the corresponding crash and exposure quantities.

We assume that, within a given month, crash incidence is proportional to exposure under a "no special risk" baseline. This yields the expected number of crashes

$$\widehat{A}_{i,t} = A_{\cdot t} \frac{E_{i,t}}{E_{\cdot t}}, \qquad A_{\cdot t} = \sum_i A_{i,t}, \;\; E_{\cdot t} = \sum_i E_{i,t},$$

where sums are taken over all segments and junctions. Segment-level and junction-level crashes are disjoint by construction, so the pooled totals $A_{\cdot t}$ and $E_{\cdot t}$ define a shared baseline.

Although crashes and exposure are aggregated at monthly resolution in preprocessing, routing requires a stable long-run risk surface. We therefore pool monthly quantities across the full study period to obtain

$$A_i = \sum_t A_{i,t}, \qquad E_i = \sum_t E_{i,t},$$

and define the corresponding baseline expectation

$$\widehat{A}_i = A_{\cdot} \frac{E_i}{E_{\cdot}}, \qquad A_{\cdot} = \sum_i A_i, \;\; E_{\cdot} = \sum_i E_i.$$

To stabilize estimation under sparse observations, we introduce a latent relative-risk multiplier θ_i and model

$$A_i \mid \theta_i \sim \text{Poisson}(\widehat{A}_i \, \theta_i), \qquad \theta_i \sim \text{Gamma}(\alpha, \alpha),$$

where the Gamma distribution is parameterized in shape–rate form, enforcing a unit prior mean $\mathbb{E}[\theta_i] = 1$. The hyperparameter α is estimated from the pooled data using a

method-of-moments estimator. By conjugacy, the posterior mean

$$r_i = \mathbb{E}[\theta_i \mid A_i, \widehat{A}_i] = \frac{A_i + \alpha}{\widehat{A}_i + \alpha}$$

serves as our Empirical Bayes smoothed relative risk. This estimator shrinks extreme values toward 1, with stronger shrinkage for entities with low expected crash counts.

Risk rescaling for routing. Relative risk estimates are dimensionless and reflect risk conditional on exposure. To obtain additive weights suitable for routing, we rescale relative risk by the pooled baseline crash rate

$$\bar{\lambda} = \frac{A_\cdot}{E_\cdot}.$$

The resulting entity-level routing weight is

$$w_i = \bar{\lambda}\, r_i,$$

which can be interpreted as an exposure-adjusted crash rate up to a global constant. This rescaling preserves relative differences in risk while yielding quantities that aggregate naturally along routes.

Routing graph. We build an undirected graph $G = (V, E)$ from the street network. Nodes represent segment endpoints and edges represent street segments with length ℓ_e. Each graph edge e corresponds to a street segment s and inherits its routing weight, denoted $w_e = w_s$. Junction identifiers and weights are mapped to nodes via spatial snapping in a projected coordinate system, yielding 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 account for both segment-level and junction-level risk, the risk contribution of edge $e = (u, v)$ is defined as

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

where w_u and w_v are junction routing weights, set to zero for non-junction nodes, and $\eta \geq 0$ controls the contribution of junction risk. We interpret these quantities as an additive surrogate for cumulative route risk, not as probabilities.

Given an origin–destination pair, the baseline route P_{dist} minimizes $L(P)$. The safety-aware route solves

$$\begin{aligned} P_{\text{safe}} = \arg\min_P R(P) = \sum_{e \in P} \rho_e \\ \text{s.t. } L(P) \leq (1+\varepsilon)\, L(P_{\text{dist}}), \end{aligned} \tag{1}$$

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

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

then select among feasible candidates the route with minimal $R(P)$, breaking ties by shorter $L(P)$. Shortest paths are computed using Dijkstra's algorithm.

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

$$\Delta_L = \frac{L(P_{\text{safe}}) - L(P_{\text{dist}})}{L(P_{\text{dist}})}$$

and the relative risk reduction

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

Pairs with $R(P_{\text{dist}}) = 0$ are excluded from Δ_R due to an undefined denominator. These metrics quantify how bounded detours trade distance for reductions in exposure-adjusted crash risk.

4. Related work.

Prior work has aimed to avoid conflating danger with demand by normalizing bicycle crashes by cyclist exposure (Lücken, 2018). City-scale studies demonstrate that exposure-normalized risk yields more informative spatial patterns than raw crash counts, and that finer temporal resolution improves inference, though persistent under-reporting in police records remains a challenge (Uijtdewilligen et al., 2024). A central obstacle in this line of research is obtaining reliable exposure: earlier methods extrapolate city-wide volumes from sparse counters using learning-based models and multi-source features, while short-term measurement campaigns improve predictions at unseen locations (Kaiser et al., 2025a). More recent efforts leverage street-segment datasets of measured bicycle volumes, enabling downstream safety analyses without the need to model exposure (Kaiser et al., 2025b). At the network level, studies define risk as crashes per unit exposure on links and address practical challenges such as spatial snapping of crash events, assigning incidents to intersections, and integrating safety metrics into routing under convenience constraints (Wage et al., 2022). The role of intersection safety is emphasized across multiple studies, which document strong crash concentrations at junctions and highlight the importance of controlling for exposure when comparing infrastructure types or locations (Medeiros et al., 2021).

Table 1. Distance–risk trade-off under bounded detours for different junction-risk weights η. Values are aggregated over all origin–destination pairs. Medians are reported with interquartile ranges in parentheses.

η	ε	Med. Δ_L	Med. Δ_R	$P(\Delta_R > 0)$
0.0	0.05	0.009 (0.026)	0.246 (0.451)	76.1
	0.10	0.025 (0.042)	0.377 (0.388)	86.0
	0.20	0.038 (0.072)	0.425 (0.353)	90.6
0.5	0.05	0.008 (0.026)	0.208 (0.401)	75.3
	0.10	0.026 (0.046)	0.331 (0.370)	86.4
	0.20	0.047 (0.089)	0.404 (0.323)	92.0
1.0	0.05	0.008 (0.026)	0.185 (0.363)	75.9
	0.10	0.028 (0.047)	0.305 (0.345)	86.3
	0.20	0.050 (0.086)	0.378 (0.318)	91.8

5. Results

To evaluate the routing algorithm, we sample $n = 1000$ origin–destination pairs uniformly at random and compare shortest-distance routes with safety-aware alternatives (Natera Orozco et al., 2020).

Table 1 summarizes the trade-off between route length and exposure-adjusted crash risk under bounded detours. Safety-aware routing identifies feasible alternatives for all origin–destination pairs across detour budgets and junction-risk weights.

Allowing a 10% detour reduces exposure-adjusted crash risk by 31–38% in median, with over 86% of routes achieving a risk reduction for all values of the junction-risk weight. Larger detours further increase these gains, reaching median reductions of 38–43% at $\varepsilon = 0.20$, while even small detours ($\varepsilon = 0.05$) yield measurable reductions of 18–25%. Across all detour budgets, increasing η is associated with lower median risk reductions.

6. Discussion and Conclusion

Our results indicate that substantial reductions in exposure-adjusted crash risk can be achieved with relatively small increases in route length. While higher junction-risk weights reduce the magnitude of the estimated risk reduction, the distance–risk trade-off persists across all configurations, with absolute gains depending on the chosen weighting.

Overall, allowing a 10% increase in route length yields median exposure-adjusted risk reductions of 31–38% for the majority of routes, demonstrating that safety-aware routing can effectively trade modest detours for meaningful safety improvements.

We provide implementation details, hyperparameters, and supplementary material, available at https://github.com/ytobiaz/data_literacy.

Contribution Statement

Explain here, in one sentence per person, what each group member contributed. For example, you could write: Max Mustermann collected and prepared data. Gabi Musterfrau and John Doe performed the data analysis. Jane Doe produced visualizations. All authors will jointly wrote the text of the report. Note that you, as a group, a collectively responsible for the report. Your contributions should be roughly equal in amount and difficulty.

References

Ehrgott, M. *Multicriteria Optimization*, volume 491 of *Lecture Notes in Economics and Mathematical Systems*. Springer, Berlin, Heidelberg, 2005. ISBN 978-3-540-21398-7.

German Federal Statistical Office. Unfallatlas. https://unfallatlas.statistikportal.de/, 2025. Interaktive Kartenanwendung zu Straßenverkehrsunfällen mit Personenschaden in Deutschland.

Hakkert, A. S., Braimaister, L., and Van Schagen, I. The uses of exposure and risk in road safety studies. Technical report, SWOV Institute for Road Safety SWOV, Leidschendam, 2002.

Kaiser, S. K. Data from: Spatio-temporal graph neural network for urban spaces: Interpolating citywide traffic volume, 2025. URL https://doi.org/10.5281/zenodo.15332147.

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:e13, 2025a. doi: 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.

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, December 2020. ISSN 2054-5703. doi: 10.1098/rsos.201130. URL http://dx.doi.org/10.1098/rsos.201130.

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.