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 sections of the street network. This makes it difficult to identify the streets that are dangerous and should be avoided and prioritised for safety improvements. We therefore use Strava's bike trip data to estimate the relative crash risk across street segments and junctions in Berlin. These risk estimates 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 prioritised for safety improvements.

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). To reveal high-risk locations, crashes must be normalised by cyclist counts; otherwise, dangerous streets can remain hidden in dense urban networks (Uijtdewilligen et al., 2024). Unfortunately, comprehensive street-level cyclist counts are rarely available. Berlin, for example, provides hourly counts at selected locations via official counting stations, but their limited spatial coverage (20 stations for thousands of streets) makes them impractical for city-wide risk estimation (Senatsverwaltung für Mobilität, Verkehr, Klimaschutz und Umwelt).

We address this problem by using bike trip counts from the fitness-tracking app Strava. These have been used to predict official counting-station data (Dadashova et al., 2020) and we show that they can serve as a proxy for cyclist exposure. For all segments and junctions in Berlin's official cycling network, we estimate exposure-normalised relative risk, defined as the ratio of observed to expected crashes. Because Strava coverage can be sparse, we use empirical Bayes smoothing for estimation (Clayton & Kaldor, 1987), which stabilises estimates in low-exposure segments and junctions, and quantifies uncertainty. Additionally, building on these estimates, we introduce a routing algorithm that finds substantially lower-risk alternatives under a route-length constraint.

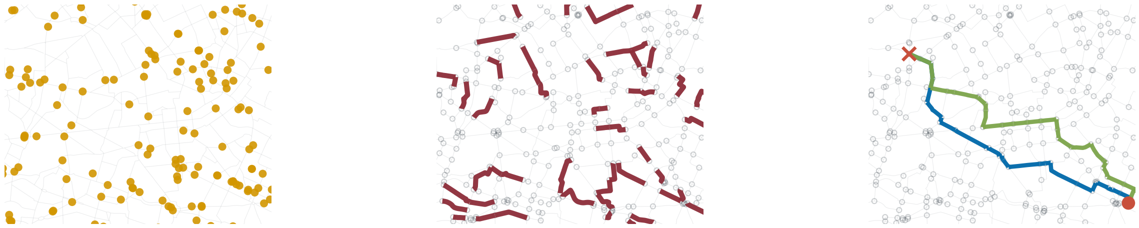
2. Data

Multiple datasets were used for risk estimation. Crash counts were taken from the *German Accident Atlas* (Destatis, 2025), which provides georeferenced locations 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 data and official counts are available (Figure 2). Count shares correlate strongly ($r = .61$) and are overall well preserved in the Strava data. Segments on wide main streets (e.g., Karl-Marx-Allee) are overrepresented in the Strava data, likely reflecting faster rides that are more often tracked, whereas residential streets (e.g., Kollwitzstraße) are underrepresented, consistent with slower, local cycling that is less often tracked.

All datasets were combined into one dataframe and matched to the same street network and map projection. The network is represented as polyline segments with associated monthly exposure counts. We map crashes to the network

[*] 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. Copyright 2025 by the author(s).



(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 crash 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 route risk under a distance-detour constraint. Filled circle and cross denote origin and destination, respectively.

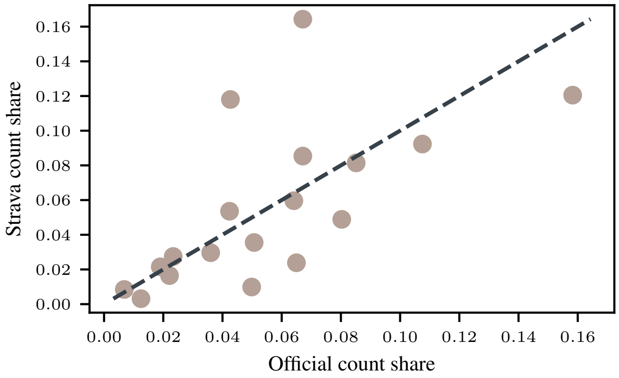


Figure 2. Consistency check between official bicycle counts and Strava-derived cyclist volumes at the street-segment level (2023). Points show segment-wise shares of total annual counts; the dashed line denotes equality between the two measures.

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. Junction exposure is derived from the segment exposure (see Section 3 for details).

At monthly resolution, events are sparse: in a typical month, fewer than 5% 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 dataset comprises 4,335 segments, 2,862 junctions, and 15,396 recorded bicycle crashes. The resulting segment- and junction-level relative risk estimates are used throughout the analysis and serve as inputs to the routing algorithm.

3. Methods

Empirical Bayes relative risk and uncertainty. For each month t, let $A_{s,t}$ and $E_{s,t}$ denote the number of police-recorded bicycle crashes and measured cyclist exposure on street segment s. Junction crashes $A_{v,t}$ 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, monthly crash incidence is assumed proportional to exposure, yielding 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 jointly over all segments and junctions, defining a shared baseline. Because routing requires a pooled baseline risk estimate, crashes and baseline expectations are aggregated over the full study period,

$$A_i = \sum_t A_{i,t}, \qquad \widehat{A}_i = \sum_t \widehat{A}_{i,t},$$

and the raw relative risk is $A_i / \widehat{A}_i$.

To stabilize estimation under sparsity, 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 prior (shape–rate) enforces $\mathbb{E}[\theta_i] = 1$ (Lord & Mannering, 2010). The hyperparameter α is estimated from the data, such that inference on θ_i is performed using Empirical Bayes smoothing.

Under the Poisson–Gamma model,

$$\mathbb{E}[A_i] = \widehat{A}_i \qquad \text{and} \qquad \mathrm{Var}(A_i) = \widehat{A}_i + \frac{\widehat{A}_i^2}{\alpha}.$$

The shrinkage parameter α is estimated via the method of moments (Morris, 1983) by equating empirical and theoretical second moments across entities,

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

Inference for entity-specific risk follows directly from Bayes' theorem,

$$p(\theta_i \mid A_i, \widehat{A}_i) \;\propto\; p(A_i \mid \theta_i, \widehat{A}_i)\, p(\theta_i),$$

which, by conjugacy of the Poisson likelihood and Gamma prior, yields the posterior distribution

$$\theta_i \mid A_i, \widehat{A}_i \sim \mathrm{Gamma}(A_i + \alpha,\ \widehat{A}_i + \alpha). \tag{1}$$

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 the Empirical Bayes relative risk, with stronger shrinkage toward the baseline for entities with low expected counts (Hauer et al., 2002). Uncertainty is quantified using $(1 - \delta)$ credible intervals obtained from quantiles of the Gamma posterior.

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

$$\bar{\lambda} = \frac{A.}{E.}, \qquad A. = \sum_i A_i, \;\; E. = \sum_i E_i.$$

The resulting routing weight is

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

We construct an undirected graph $G = (V, E)$ from the street network. Nodes correspond to segment endpoints and edges to street segments with length ℓ_e. Each edge e corresponds to a segment s and inherits its routing weight $w_e = w_s$. 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 segment- and junction-level risk, the risk contribution of edge $e = (u, v)$ is

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

where w_u and w_v are junction routing weights (zero for non-junction nodes) and $\eta \geq 0$ controls the contribution of junction risk. These quantities form an *additive surrogate* for cumulative route risk.

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

$$\begin{aligned} P_{\text{safe}} = \arg\min_P R(P) &= \sum_{e \in P} \rho_e \\ \text{s.t. } L(P) &\leq (1 + \varepsilon)\, L(P_{\text{dist}}), \end{aligned} \tag{2}$$

where ε is the allowable relative detour (Ehrgott, 2005). We approximate this constraint via 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 among feasible candidates the route minimizing $R(P)$. Shortest paths are computed with Dijkstra's algorithm (Dijkstra, 1959).

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 the undefined denominator. These metrics quantify the trade-off between distance and exposure-adjusted crash risk under bounded detours.

4. Related work.

Prior work seeks to avoid conflating danger with demand by normalizing bicycle crashes by cyclist exposure (Lücken, 2018). City-scale studies show 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 is obtaining reliable exposure: earlier work extrapolates 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 instead rely on street-segment datasets of measured bicycle volumes, enabling safety analyses without explicit exposure modeling (Kaiser et al., 2025b). At the network level, risk is typically defined as crashes per unit exposure on links, with attention to spatial snapping, assignment of incidents to intersections, and integration of safety metrics into routing under convenience constraints (Wage et al., 2022). Intersection safety is repeatedly emphasized, with strong crash concentrations at junctions and the need to control for exposure when comparing locations or infrastructure types (Medeiros et al., 2021).

5. Results

One of the major outcomes of this work ist the estimation of the risk to have get injured in a bike accidents while driving on a streetsegment. This distribution of those calculated risks is displayed in .

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. 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. Δ_L and Δ_R are reported on a relative scale, whereas $P(\Delta_R > 0)$ is reported in percent.

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

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.

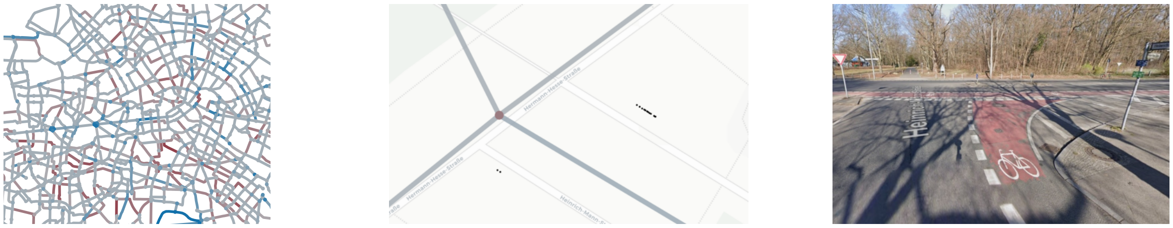
Allowing a 10% detour reduces exposure-adjusted crash risk by 31–38% in median, with over 86% of routes achieving a risk reduction for all values of the junction-risk weight. Larger detours further increase these gains, reaching median reductions of 38–43% at $\varepsilon = 0.20$, while even small detours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



(a) Risk heatmap

(b) High-risk junction

(c) Street view

Figure 3. **Risk patterns and detailed inspection of a high-risk junction in Berlin.** (a) Network-wide cyclist risk map, where colors (▬) indicate log-scaled relative risk from low (blue) to high (red). (b) Detail view of junction 2482 (Hermann-Hesse-Straße / Heinrich-Mann-Straße) with recorded accident locations. A total of 22 accidents occurred at this junction, predominantly involving a second vehicle, mostly cars (20 cases), with two involving other vehicle types. Most accidents were associated with vehicles turning into or crossing the road (20), while only two resulted from turning off the road. (c) Street-level view of the junction (Google, 2025), illustrating that the car lane crosses the bicycle lane.

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

Clayton, D. and Kaldor, J. Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, pp. 671–681, 1987.

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*, 2674(11):390–402, 2020.

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, December 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.

Hauer, E., Harwood, D. W., Council, F. M., and Griffith, M. S. Estimating safety by the empirical bayes method: A tutorial. *Transportation Research Record: Journal of the Transportation Research Board*, 1784(1):126–131, 2002. ISSN 2169-4052. doi: 10.3141/1784-16. URL http://dx.doi.org/10.3141/1784-16.

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.

Lord, D. and Mannering, F. The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives. *Transportation Research Part A: Policy and Practice*, 44(5):291–305, 2010. ISSN 0965-8564. doi: 10.1016/j.tra.2010.02.001. URL http://dx.doi.org/10.1016/j.tra.2010.02.001.

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. Radverkehrszählstellen – jahresbericht 2023, May 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.

Senatsverwaltung für Mobilität, Verkehr, Klimaschutz und Umwelt. Zählstellen und fahrradbarometer: Fahrradverkehr in zahlen. URL https://www.berlin.de/sen/uvk/mobilitaet-und-verkehr/verkehrsplanung/radverkehr/weitere-radinfrastruktur/zaehlstellen-und-fahrradbarometer/.

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.