On cycling risk and discomfort: urban safety mapping and bike route recommendations

David Castells-Graells[1] · Christopher Salahub[2] · Evangelos Pournaras[3]

Abstract

Bike usage in Smart Cities is paramount for sustainable urban development: cycling promotes healthier lifestyles, lowers energy consumption, lowers carbon emissions, and reduces urban traffic. However, the expansion and increased use of bike infrastructure has been accompanied by a glut of bike accidents, a trend jeopardizing the urban bike movement. This paper leverages data from a diverse spectrum of sources to characterise geolocated bike accident severity and, ultimately, study cycling risk and discomfort. Kernel density estimation generates a continuous, empirical, spatial risk estimate which is mapped in a case study of Zürich city. The roles of weather, time, accident type, and severity are illustrated. A predominance of self-caused accidents motivates an open-source software artifact for personalized route recommendations. This software is used to collect open baseline route data that are compared with alternative routes minimizing risk and discomfort. These contributions have the potential to provide invaluable infrastructure improvement insights to urban planners, and may also improve the awareness of risk in the urban environment among experienced and novice cyclists alike.

David Castells-Graells and Christopher Salahub contributed equally to this work.

✉ Evangelos Pournaras
  e.pournaras@leeds.ac.uk

  David Castells-Graells
  david.castells@mpq.mpg.de

  Christopher Salahub
  csalahub@uwaterloo.ca

1   Max-Planck-Institut für Quantenoptik, Hans-Kopfermann-Str. 1, 85748 Garching, Germany

2   Department of Statistics and Actuarial Science, University of Waterloo,
    200 University Avenue West, Waterloo N2L 3G1, Canada

3   School of Computing, University of Leeds, Leeds LS2 9JT, UK

🖄 Springer

1 Introduction

Bikes are transforming urban environments in Smart Cities [2,12] and advancing challenging sustainable development goals: cycling supports healthier lifestyles, decreases energy consumption and carbon emissions in urban centers, and reduces traffic jams. Furthermore, cycling reduces the need for extensive car parking infrastructure, providing opportunities for the development of parks, greenery, and recreation areas. Bikes provide benefits not only for residents, but also give tourists an engaging means to explore a city. The massive expansion of bike sharing infrastructure in urban centers is demonstrative of this transformation.

Positive as they are, these developments have come with an increase in the rate of bike accidents. In Zürich, for example, while cycling traffic has increased by 35% since 2013, reported bike accidents have increased[1] by 60%. The accident rate of cyclists is 5 to 6 times higher per travelled kilometer than that of car occupants [9] in Norway. When unreported bike accidents are considered, the risk to cyclists may be as much as 20 times higher than that of car occupants [9]. Additionally, almost half of a million cyclists die every year in traffic accidents [11]. New insights into the causes of bike accidents and their prevention are imperative if the bike is to be a predominant means of transportation in sustainable Smart Cities.

This paper introduces a data-driven method to estimate cyclist risk and discomfort and applies it to the city of Zürich. The proposed estimation model uses kernel density estimation to provide a *continuous estimation of risk on the traffic network*. The severity of accidents, their causes, the role of the weather and seasons, and the impact of the day and time are all extensively studied using a diverse spectrum of data sources including public authorities, health insurance policies, OpenStreetMaps (OSM) traces, and typical routes collected from Zürich cyclists. The dominance of self-caused accidents indicates the potential to improve bike safety via personalized route recommendations balancing safety and comfort generated by an open-source software artifact.

Some work on bike safety has focused on environmental and demographic factors related to cycling safety, e.g. age, gender, daylight conditions and use of helmets [13,14]. Findings include insights on the limited protection provided by helmets and the necessity that a child is developmentally ready for cycling. Data are mainly analysed at an aggregate national level in the United States and location-specific risks are not taken into account.

Other work has produced risk metrics such as *exposure* that account for the distance cycled before an accident occurs [10]. Measuring exposure requires the choice of specific areas and times for modeling. In contrast, the approach introduced in this paper generalizes risk estimation to a continuous geographic area. Moreover, the concept of exposure conveys only the potential of an accident, while the risk estimate of this paper is explicitly based on actual accident data reported to official authorities. The relation

---

[1] Available at https://www.limmattalerzeitung.ch/limmattal/zuerich/autofahren-wird-sicherer-auf-zwei-fraedern-ist-das-gegenteil-der-fall-132314683 (last accessed: May 2019).

between risk and exposure as well as other methodologies to model and measure cycling safety are reviewed extensively in a recent report by the U.S. Department of Transportation [17], though this report does not provide any quantitative data analysis as performed in this paper.

More relevant related work assesses the risk of cycling routes in Berlin by counting "hot spots" (high risk areas) and dangerous intersections within routes [7]. No continuous risk estimation is provided and all route features are treated equally. Spatio-temporal analysis frameworks of bike risk have been introduced using network-based kernel density estimation [3,9]. For instance, applying this method in the city center of Vienna, Austria, reveals that bike accident hot spots vary in space according to season, light, and precipitation conditions. Additionally, these hot spots cluster by intersections and bus/tram/subway/bike stations [9]. Although the scope of this work is the closest to this paper, it addresses neither discomfort nor route recommendations.

The contributions of this paper are as follows: (i) A continuous spatial risk and route discomfort estimation model for mapping cyclists' urban safety. (ii) A personalized route recommendation system that balances cycling risk and discomfort. (iii) The design of a novel data analytics pipeline that combines input from a diverse spectrum of data sources to map cycling risk. (iv) Findings about the number of accidents, their severity and their causes in the area of Zürich; the influence of weather/seasonality as well as daily/weekly accident patterns. (v) An open-source software artifact[2] for the interactive collection of bike route data as well as the computation of personalized route recommendations. (vi) An open dataset [15] of cyclists' bike routes in the area of Zürich that can be used as a baseline for multi-objective bike route optimization.

This paper is organized in the following sections: Sections 2 and 3 introduce the spatial risk and route discomfort estimation models respectively. Section 4 introduces the concept of personalized route recommendations that balance safety and comfort. Section 5 illustrates the experimental methodology for the evaluation of the risk estimation model as well as a software artifact for data collection and bike route recommendations. Section 6 presents the findings of the performed data analysis. Finally, Sect. 7 concludes this paper and outlines future work.

2 A spatial risk estimation model

This section introduces a general-purpose data-driven model for the spatial estimation of transport risk using geolocated traffic and accident data. Table 1 illustrates the main mathematical symbols and their data applicability. Risk is the conditional probability of involvement in a traffic accident A given the use of a particular transit method T. It is represented by a conditional probability density $A|T$ as follows:

$$f_{A|T}(\mathbf{x}) = \frac{f_{A,T}(\mathbf{x})}{f_T(\mathbf{x})}, \tag{1}$$

where $f_{A,T}(\mathbf{x})$ is the joint density and $f_T(\mathbf{x})$, $f_A(\mathbf{x})$ are the appropriate marginal densities. $f_{A,T}(\mathbf{x})$ can be viewed as a normalization of $f_A(\mathbf{x})$ to account for the level

---

Table 1 Notation of the spatial risk estimation model and its applicability

Notation	Meaning	Applicability
$\mathbf{x} \in \Re^m$	Estimation coordinate	Zürich street network grid
$\mathbf{x}_i \in \Re^m$	Observation coordinate	Accident and traffic data
m	Dimension of coordinates	2
n	Number of observations	1305 Accidents, 242801 GPS trace points
h	Scalar bandwidth parameter	0.003
$K_h(\cdot)$	h-Parameterized kernel	$\frac{1}{\sqrt{2\pi h}^m} e^{\frac{\mathbf{x}^T \mathbf{x}}{2h}}$
$\mathbf{x}^T$	Transpose operation	–
$\hat{f}$	Estimator of f	–
A_s	sth Partition of the accident data	Accidents with severity s
S	Number of partitions	3

of traffic such that a street with 4 accidents over 15 trips has a lower risk than a street with 3 accidents over 10 trips, for example. In practice, the estimation of the conditional density, $f_{A|T}(\mathbf{x})$, is feasible with geolocated accident and transit data for the transit method T, representing samples from the joint distribution $f_{A,T}(\mathbf{x})$ and the marginal density $f_T(\mathbf{x})$ respectively. Given that the individual accident coordinates $\{\mathbf{x}_i\}_{i=1,\dots,n}$ are discrete points, a continuous and non-parametric density estimate can be calculated using kernel density estimation (KDE) [6].

Given n points in m dimensions and a local density function, or kernel, the equation for estimating the density at point $\mathbf{x}$ using the set $\{\mathbf{x}_i\}_{i=1,\dots,n}$ for a given kernel K parameterized locally by h is given as follows:

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_h(\mathbf{x} - \mathbf{x}_i),$$

(2)

where K_h is the kernel function normalized such that integration over its local support is one, e.g. the Gaussian density:

$$K_h(\mathbf{x}) = \frac{1}{(\sqrt{2\pi h})^m} e^{\frac{\mathbf{x}^T \mathbf{x}}{2h}}.$$

(3)

An isotropic zero-centered Gaussian kernel ensures that the local density estimation does not preferentially estimate densities in any transit direction, a desirable feature if the orientation of streets within a region is not known a priori. However, this can spuriously relate transit points, e.g. accidents that occur on parallel but disconnected streets.

Kernel density estimation gives equal influence to each geolocated accident used in estimation. However, additional meta-information can be employed to distinguish certain accident locations, for instance, and assign weights to accidents by severity. Assume that $\{\mathbf{x}_i\}_{i=1,\dots,n}$ can be partitioned into S meaningful subsets indexed by s.

Classifying the geolocated accident data by severity gives subsets A_s of size $n_s = |A_s|$ labeled by severity meta-information. The density is reestimated in this case according to Corollary 1.

Corollary 1 *The kernel density estimate of geolocated accident data classified in S subsets each with size n_s is calculated as follows:*

$$\hat{f}(\boldsymbol{x}) = \sum_{s=1}^{S} \frac{n_s}{n} \hat{f}_{A_s}(\boldsymbol{x}), \tag{4}$$

where $\hat{f}_{A_s}$ is the kernel density estimate of the data in subset A_s.

Proof The sum over all n elements can be expanded as the sum over all subsets and subset elements $\mathbf{x}_{sj}$ as follows:

$$\hat{f}(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} K_h (\mathbf{x} - \mathbf{x}_i) = \frac{1}{n} \sum_{s=1}^{S} \sum_{j=1}^{n_s} K_h (\mathbf{x} - \mathbf{x}_{sj}). \tag{5}$$

The summands can be multiplied with $1 = \frac{n_s}{n_s}$. However, n_s is constant within the second summand over A_s and as such the numerator is moved outside the inner sum. Similarly, $\frac{1}{n}$ can be moved inside the first summand. Based on Eq. 3 the kernel density estimation of partition s is derived:

$$\frac{1}{n} \sum_{s=1}^{S} \sum_{j=1}^{n_s} K_h (\mathbf{x} - \mathbf{x}_{sj}) = \sum_{s=1}^{S} \frac{n_s}{n} \sum_{j=1}^{n_s} \frac{1}{n_s} K_h (\mathbf{x} - \mathbf{x}_{sj}) = \sum_{s=1}^{S} \frac{n_s}{n} \hat{f}_{A_s}(\mathbf{x}). \tag{6}$$

□

Partitions can be reweighted to reflect the typical consequence of an accident in a subset, for instance insurance compensations for accidents of the S partitions, or the relative importance of the partitions S to a policy directive. The following equation:

$$f_R(\mathbf{x}) = \sum_{s=1}^{S} a_s \hat{f}_{A_s}(\mathbf{x}), \tag{7}$$

is such a generalization, where $\frac{n_s}{n}$ corresponds to weights a_s such that $\sum_{s=1}^{S} a_s = 1$.

3 A route discomfort estimation model

Along with risk estimation, discomfort estimation can be used to assign a weight to each street of a street network. The *discomfort* of a bike route is defined by the level of physical effort required to traverse the route by bike in terms of length and grade. Characterising the contributions of the two variables with a closed expression is not straightforward. A vast amount of data and domain experience is required to design

a realistic model. The *IBP index*[3] reflects such a model of discomfort. It is generated by an algorithm that utilises years of iteration between measurement and adjustment to assess the cycling difficulty of a bike route. This index is currently used by many associations and guides such as the French Federation of Hiking. Although a series of corrections and fits accounting for multiple parameters mean that the IBP index is not based on a single closed expression, it can be used to calibrate one such expression.

To make use of the the IBP index, different synthetic tracks of constant grade and specific length are generated and provided to the IBP index application in the required GPX format. A linear dependency with length and an exponential one with grade is inferred after inspecting the outputs and the fitting curves. The discomfort expression is given in the following form:

$$f(d, x) = \begin{cases} d \cdot (2 \exp(15\,x) - 1) & \text{if } x \geq -0.025 \\ f(d, -0.025) & \text{otherwise,} \end{cases} \tag{8}$$

where d is the length of a street or route and x is the average grade: $-1 \leq x \leq 1$. The expression is set constant below a -2.5% grade, as grade values lower than this threshold make no difference in effort for a cyclist. This limit agrees with results from the IBP index calculator. Note that at zero grade, $f(d, 0) \propto d$.

4 Personalized route reccomendation based on risk and discomfort

Personalized route recommendations for given departure and destination points are calculated using Breadth First Search (BFS) over the street network with assigned weights on the streets. These weights are a cyclist's determined combination of the risk and discomfort estimates as follows:

$$w = \alpha\,w_{\mathrm{r}} + (1 - \alpha)\,w_{\mathrm{d}}, \tag{9}$$

where w_{r}, w_{d} are the risk and discomfort weights respectively. An $\alpha = 0$ prioritizes routes with minimal discomfort, while $\alpha = 1$ prioritizes routes with minimal risk.

5 Data science and experimental methodology

This section introduces a realization of the proposed risk estimation model and a software artifact for bike route recommendations. Figure 1 outlines the process of bike risk assessment in the city of Zürich. A data science pipeline is designed that consists of the following stages: (1) Extraction and categorization of bike accident data in JSON format from the Swiss GeoAdmin[4] API[5] to obtain $\{\mathbf{x}_i\}_{i=1,\dots,n} = \{A_s\}_{s=1,\dots,S}$. (2)

---

[3] Available at https://www.ibpindex.com (last accessed: May 2019).

[4] Available at https://map.geo.admin.ch/?topic=vu&lang=de&bgLayer=ch.swisstopo.pixelkarte-grau&layers=ch.astra.unfaelle-personenschaeden_alle&layers_timestamp=&catalogNodes=1318 (last accessed: May 2019).

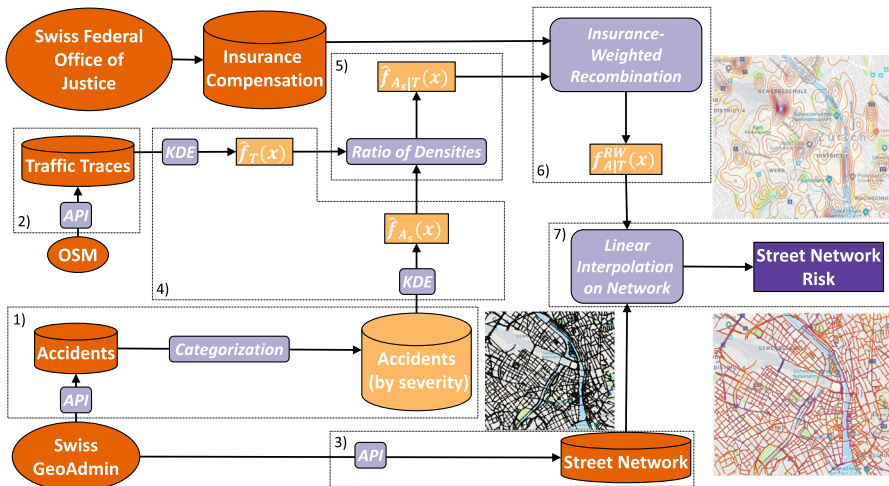[5] Available at http://api.geo.admin.ch (last accessed: May 2019).

Fig. 1 An overview of the data science pipeline for bike riding risk assessment

Extraction and processing of GPS trip traces in XML format from the OpenStreetMaps (OSM) API.[6] (3) Extraction and processing of the Zürich street network in JSON format from the Swiss GeoAdmin[5] API.[6] (4) Kernel density estimation on traces and labeled accidents to calculate $f_T(\mathbf{x})$ and $f_{A_s,T}(\mathbf{x}) \forall s$. (5) Calculation of $\hat{f}_{A_s|T}(\mathbf{x}) \forall s$ according to Eq. 1 by taking the ratio of $\hat{f}_T(\mathbf{x})$ and $\hat{f}_{A_s,T}(\mathbf{x}) \forall s$. (6) Application of insurance recompensation data[7] from the Swiss Federal Office of Justice to obtain $f_R(\mathbf{x})$ from $\hat{f}_{A_s|T}(\mathbf{x})$ according to Eq. 7. (7) Interpolation of $f_R(\mathbf{x})$ onto the processed street network.

5.1 Stage 1: Accident data extraction

The Swiss Confederation[5] web portal has an interactive map of Switzerland with several spatial layers of publicly available data. One of these layers compiles and displays accidents involving bikes between 2011 and 2017. The data are collected by the Swiss Federal Roads Office from electronic police reports.[8] Together with the localization of the accidents, this layer provides the date, time, severity, cause, and street type of the accident. These features are visualized on the map, which also serves as the basis of an API service for batch data extraction.[6]

The API service 'identify' is used for data extraction. It generates and returns a list of at most 200 elements from a layer, e.g. individual accidents from the accident layer or road network points from the road network layer, satisfying a geometry specified using

---

(a) Selected data region.   (b) Iteration 1   (c) Iteration 2   (d) Iteration 3   (e) Iteration 4
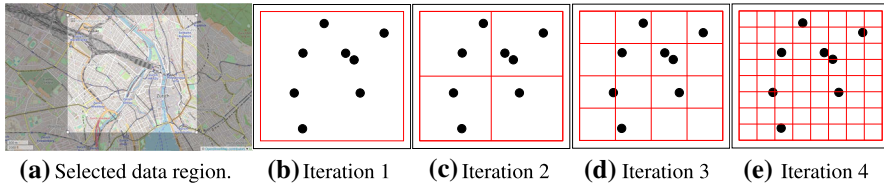
Fig. 2 The selected region and an example of subdivisions using a threshold of 1

ESRI syntax.[9] For simplicity and scalability, the geometry used in this investigation is a bounding box specified by its horizontal and vertical extents as shown in Fig. 2a. It is delimited[10] by the latitudes (47.3650, 47.3886) and longitudes (8.5141, 8.5523). This region is chosen for its central location in Zürich and the density of recorded trips on OSM, an indication of high traffic volumes.

To conform to the limit of 200 extracted elements, consecutive subdivisions of the bounding box are performed prior to data extraction so that each subdivision in which the final data extraction[11] takes place contains 200 elements or fewer. Figure 2b– e displays this algorithm schematically using a threshold of 1 rather than 200 for simplicity and Algorithm 1 outlines the subdivision logic.

Algorithm 1 Grid subdivision logic of the selected region.

---
set grid subdivision to 1
while number of elements in any subdivision > 200 **do**
    divide each subdivision into quarters
end while
return Subdivision boundaries

The 'severity' field from the extracted data is used to categorize the different accidents into those resulting in *light* injuries, *severe* injuries, and *death*.

5.2 Stage 2: Trip data extraction

Transit data from users' GPS traces are downloaded in XML format from OSM[7] and treated as a sample of $f_T(\mathbf{x})$. 5% of the traces are labeled by means of transport, from which all non-bike traces are removed to improve the quality of $f_T(\mathbf{x})$ estimation. For unlabeled traces, travel homogeneity between methods of transit is assumed as streets included in the selected data window are primarily multi-use.

---

[9] This data are defined using the Swiss projection coordinates of the Swiss GeoAdmin system: https://www. swisstopo.admin.ch/en/maps-data-online/calculation-services.html (last accessed: May 2019). All values are converted to the more common WGS84 system using the approximation.

[10] The 'identify' service requires the specification of other map features that affect the inclusion tolerance of any bounding box: (i) the *map extent* parameter that is the entirety of Switzerland set with 312,250, −77,500, 1,007,750, 457,500 and (ii) the *image display* parameter chosen to be 1391, 1070, 96.

[11] To ensure all data points are obtained including those on the boundaries, the final extraction windows applied to each subdivision are selected with deliberate overlap at the boundaries. This process creates duplicate records that are removed using the unique ID associated with each element.

This assumption is also motivated by the very limited selection of data that can be used in the scope of this paper. For instance, Open Data Zürich[12] has more precise bike traffic data, but they lack the resolution to be used for kernel density estimation. Though this assumption imposes limitations on the generalization of the performed analysis and may introduce imprecision, its impact on analysis did not seem to be critical, as discussed in Sect. 5.4.

5.3 Stage 3: Street network data extraction

The Zürich street network required for $f_{A|T}(\mathbf{x})$ is extracted from the Swiss Confederation web portal[13] as shown in Stage 1. It is specified by the coordinates (latitude, longitude, altitude) of points along the city streets and the street segments linking these points. While the street points are not equidistant, they are always present at intersections. The large number of points defining the street network makes computations, such as the graph search used for route recommendations, very expensive. The computational load is reduced by keeping the intersection points only, which is roughly 10% of the total.

After the data extraction and pre-processing, the extracted data are modeled as a graph structure. Each street network point is assigned to a node and the points are connected by the edges representing street segments. The process is quite accurate, as the source data are already grouped into street segments with matching coordinates at the intersections, with some tolerance in the last decimal digits.

5.4 Stage 4–7: Risk estimation

The kernel density estimation is performed using the `kde2d` function in the `stats` package[14] of R. As suggested in Eq. 2, different choices of h result in very different estimated densities. A large h gives a coarse and uninformative density, while a small h results in a density that cannot effectively estimate beyond the immediate neighborhood of each $\mathbf{x}_i$. This is visually presented in the two examples of Fig. 3a, b. A bandwidth of 0.003°, in WGS84 coordinates, is selected empirically as it provides a good trade-off. The estimated density contours across the entire accident data are shown in Fig. 3c.

The density $f_{A_s,T}(\mathbf{x})$ of bike accidents is calculated for each severity level s. Similarly, $f_T(\mathbf{x})$ is estimated using the OSM traces to calculate Eq. 4. As the studied area is not perfectly square, a grid with 560 horizontal and 440 vertical divisions is imposed and estimations are made at the intersections of the grid lines. This asymmetry ensures that the evaluation positions are equispaced in terms of WGS84 coordinates of latitude and longitude.

To generate $f_R(\mathbf{x})$, a relative weighting of 1:6:6 for light injuries, severe injuries, and death respectively is used to recombine the partition densities. These weights are

---

(a) Bandwidth=0.001   **(b)** Bandwidth=0.01   **(c)** Bandwidth=0.003
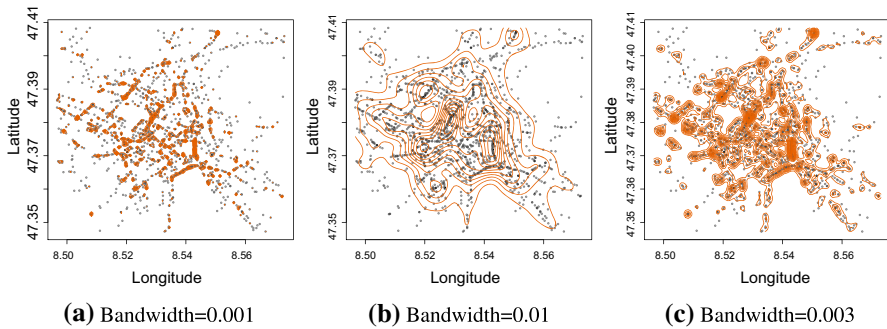
Fig. 3 Bicycle accident density contours for different bandwidth choices

(a) Non-normalized contours.   **(b)** Normalized contours.   **(c)** Non-scaled interpolation.   **(d)** Scaled interpolation.
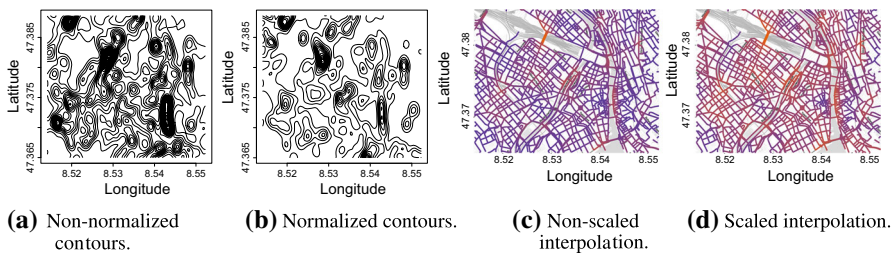
Fig. 4 Density contours and interpolated network risk in Zürich. Orange hue denotes higher risk (color figure online)

based on: (i) Insurance compensation policy data of 5000:30,000:100,000 CHF for the respective accident severity levels.[8] (ii) Deaths are treated as severe injuries due to their low number (5) and minor differences from severe injuries as addressed in Sect. 6.1.

It is noted in Sect. 5.2 that the assumption of travel homogeneity in the OSM traces may not always hold, leading to imprecision in $\hat{f}_T(\mathbf{x})$. This could cause spurious contour peaks despite similar accident rates, i.e. the ratio of $\hat{f}_{A_s,T}(\mathbf{x})/\hat{f}_T(\mathbf{x})$ from Eq. 1, for these areas. Non-normalized and normalized density contours are compared in Fig. 4a, b. The contour peaks of the latter are less extreme than those in Fig. 4a, while the dominant peaks remain in the same locations and distinguishable in both versions, suggesting that normalization has scaled the risk measurement of high traffic regions as desired. The similar placement of peaks suggests that the potential imprecision in OSM data has not drastically changed the patterns in the bike accident data.

Note that the estimation window in Fig. 3 extends beyond the specified studied region. Density estimation has highly variable boundary behavior due to the abrupt exclusion of points at the window edges. This boundary effect, further exacerbated by taking the ratio of densities estimated over the window, results in spuriously peaked boundary estimates of $f_{A_s|T}(\mathbf{x})$. An extended window is introduced to estimate the densities, before restricting back and normalizing to the studied region.

At the final stage, $f_R(\mathbf{x})$ is mapped to the street network using simple linear interpolation. The resulting normalized risk is plotted on a map of Zürich using the ggmap [8]

and `ggplot2` [18] packages in `R`. The interpolated risks on the street network are displayed in Fig. 4c.

Immediately apparent is the relatively high risk in two vibrantly orange areas near Hardbrücke[15] and Langstrasse.[16] These areas are, by a wide margin, the most dangerous in Zürich and the magnitude of their risk makes visual risk inspection of the rest of Zürich challenging. A Box-Cox power transformation [5] with an exponent of $\frac{1}{2}$ is applied to the data as shown in Fig. 4d. The variation in risk is more visually apparent and so it is easier to distinguish higher and lower risk areas.

The risk estimation method illustrated in this paper relies on the quality of the reported accident data. However, it is likely that accidents are under-reported to police, especially those that do not result in injuries or property damage. The following reasoning is made about these unreported accidents: (i) As unreported accidents are expected to be of light severity, they are not expected to significantly increase the estimated risk values. Moreover, cyclists are likely more interested in those accidents with recorded injuries. (ii) Under the assumption that unreported accidents are homogeneously distributed in the studied area, their influence on the estimated density contours is negligible.

5.5 A software artifact for personalized bike route recommendations

A software artifact is introduced to calculate route recommendations on a weighted graph extracted from the street network. This software accepts departure and destination points from a user and a preference for the balance of safety and comfort expressed as a weight, α, and produces a recommended route based on this information. Input is collected using an interface with an interactive map implemented in the Python *tkinter* library. The user clicks on the map to provide the departure and destination points, as well as intermediate route points, whose latitude and longitude coordinates are computed in the background. The recommended route for a given α is then displayed to the user on the map. The user input and calculated values, i.e. total risk, discomfort and points that form the route, can be exported in a `.txt` file.

The map first matches the departure and destination points to nodes of the street graph. The matching is performed by the distance minimization between the points and nodes. The graph is implemented as a Python object and the software executes the BFT algorithm[17] given the destination, departure, and α as illustrated in Sect. 4.

The software artifact and route recommendations are evaluated using two sets of data: (i) 24 typical bike routes by 8 individuals who cycle regularly in the studied area. Data collection is performed via the software artifact by clicking the interactive map several times to form a route exported as described earlier. These routes are referred to as *baseline routes* and they are compared to the route recommendations for the

---

[15] Available at https://en.wikipedia.org/wiki/Hardbrücke (last accessed: May 2019).

[16] Available at https://en.wikipedia.org/wiki/Langstrasse (last accessed: May 2019).

[17] For efficiency, the different explored paths have the accumulated weight assigned and are terminated whenever their accumulated weight exceeds the value of a path that has already reached the destination. They are also terminated whenever they reach a node previously visited by any other path, whose weight at that node is lower than the one of the current path. Once the minimal path is found, the list of nodes is returned together with the calculation of the accumulated risk and discomfort weights.
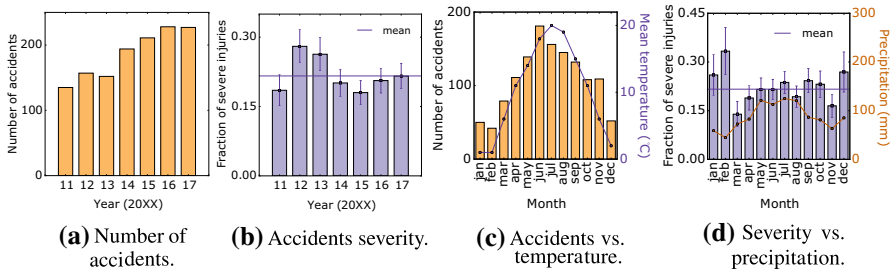
(a) Number of accidents. **(b)** Accidents severity. **(c)** Accidents vs. temperature. **(d)** Severity vs. precipitation.

Fig. 5 Yearly and monthly analysis of accident and weather data

same departure and destination points and different α values in order to assess the improvement in safety that can be gleaned by the route recommendation software. (ii) 2000 random departure and destination points are generated on the map. For each pair, three route recommendations are generated for $\alpha = 0$, $\alpha = 0.5$ and $\alpha = 0.75$. The recommended routes are then displayed on a map of Zürich, thereby facilitating a comparison of the utilization of the street network for different weights of safety to discomfort. In this way, the overall risk and discomfort estimation that stems from the route recommendations is mapped on the studied region.

6 Experimental evaluation

This section analyses the accidents and evaluates the bike route recommendations.

6.1 Accident analysis

The total number of reported accidents in the specified region and time frame is 1305: 1023 light injuries, 277 severe injuries, and 5 deaths. Because of the low number of death events, they are included in the group of severe injuries. Additionally, the occurrence of all deaths in regions with a large number of severe injuries and not elsewhere suggests that the circumstances of both types of accidents are likely very similar. Figure 5a illustrates the yearly evolution of accidents. From 2013 to 2017, an increase of approximately 50% is observed. Figure 5b shows the probability of accidents resulting in severe injuries. The values remain within the expected variation.[18]

The total number of accidents per month across all years vs. the average temperature per month[19] is shown in Fig. 5c. That fewer accidents occur in colder months reflects the temporal transit patterns of citizens in Zürich: choosing public transport or driving by car during the colder months of the year to avoid the discomfort of cycling in

---

[18] The error bars are determined by assuming a binary (Bernoulli), distribution between light and severe injuries. For the probability of a severe injury at a specific year, p, the data-based value is used as a good approximation to the true value. Therefore, the standard deviation of the mean has the following form: $\sqrt{p(1-p)/N}$, where N is the total number of accidents for each year.

[19] Available at https://www.timeanddate.com/weather/switzerland/zurich/climate (last accessed: May 2019).

(a) Accidents per type.

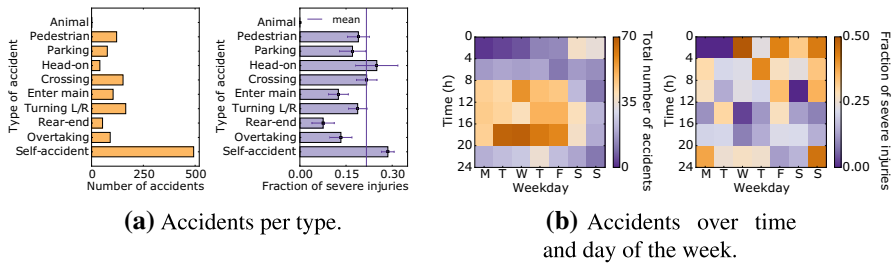**(b)** Accidents over time and day of the week.

Fig. 6 Type of accidents and their time of occurrence

the cold. On the other hand, the relative severity of the accidents is higher during the winter months, as shown in Fig. 5d. Given the often steep grades of streets in the studied region, snow and frozen street surfaces in winter are the most likely explanation of this observation. Precipitation also seems important. Summer shows on average up to 70% higher precipitation[20] than the months of March and November, during which the lowest fractions of severe injuries are observed.

Figure 6a illustrates the relation between the accidents and their causes. Self-caused accidents are predominant: 40% of all accidents are self-caused. Additionally, a relatively high proportion of these are severe, 28%, as shown in Fig. 6a. Head-on collisions and accidents on crossing lanes follow in severity, suggesting that intersections entail higher overall risk. Given the predominance of self-caused accidents, the potential to improve bike riding safety via warnings and route recommendations is apparent. Such risk communication can improve cyclist awareness and simultaneously provide greater confidence to tourists and new cyclists.

Figure 6b illustrates the time of accident occurrences during weekdays, which show significantly more accidents than weekends. Weekday accidents usually happen early morning and late afternoon, suggesting that accidents happen during commuting times, i.e. home-work and vice versa. During weekends accidents mainly occur during the following times: (i) Saturday afternoon, probably corresponding to shopping/outings. (ii) Early morning hours on Saturday and Sunday, suggesting accidents related to poor visibility conditions, fatigue, and alcohol consumption. Figure 6b further indicates that the latter are the most severe ones.

6.2 Bike route recommendations: safety versus discomfort

Figure 7 shows the relative improvement of risk and discomfort between the 24 baseline routes collected from cyclists in Zürich and the recommended routes for different α values. As expected, increasing α results in a monotonic increase in discomfort and decrease in risk. While there appears to be minimal reduction in risk and a large increase in discomfort for $\alpha > 0.4$, moving from an α value near 0 to one near 0.4 reduces risk by a greater proportion than discomfort is increased, suggesting optimal values of α lie between 0.2 and 0.4. Larger values of α are likely to be highly undesirable to cyclists, as the high discomfort suggests long routes with steep grades.

Fig. 7 Average relative
improvement of risk and
discomfort between baseline and
recommended routes. The light
purple line is the mean
improvement between the two
estimates, which can be used to
assess α values with a good
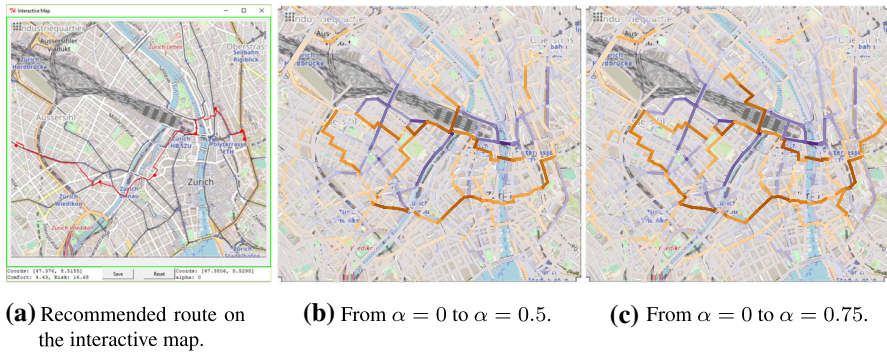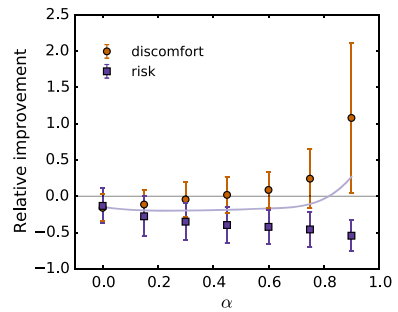balance between the two (color
figure online)

(a) Recommended route on the interactive map.

(b) From $\alpha = 0$ to $\alpha = 0.5$.

(c) From $\alpha = 0$ to $\alpha = 0.75$.

Fig. 8 The graphical user interface for interactive route recommendations and the changes in the frequency of street utilization by the recommended routes when changing from $\alpha = 0$ to $\alpha = 0.5$ and $\alpha = 0.75$ respectively. Orange street segments indicate an increase in street utilization, purple ones a decrease (color figure online)

Figure 8a displays the graphical user interface of the software artifact introduced in Sect. 5.5. The results of the random route generation discussed in Sect. 5.5 are presented in Fig. 8b, c. These figures illustrate the changes in street utilization that result from increasing α from 0 to 0.5 and 0.75, respectively. In other words, they show the changes in street use that come when a higher priority on safety is given to BFS. The 2000 randomly generated departure and destination points are used for the mapping of street utilizations.

The colored maps show that areas such as Langstrasse[17] and Bahnhofstrasse[20] are already avoided at $\alpha = 0.5$, while two cross-city routes become dominant.

7 Conclusion and future work

A data-driven approach for the estimation and mapping of cycling risk in complex evolving urban environments can provide invaluable empirical insights about safety. This is shown for the city center of Zürich, in which continuous risk contours are calculated based on historical geolocated accident data and information about their severity, linked to health insurance compensation policies. Findings show that bike

---

[20] Available at https://en.wikipedia.org/wiki/Bahnhofstrasse (last accessed: May 2019).

accidents increase at a higher rate than bike use, while weather, seasonality, day of the week, and time play a role on the likelihood of an accident and its severity. The predominance of self-caused accidents suggests the need for a greater awareness of risks and safe routing information. This requirement is met by personalized route recommendations that balance safety and comfort. The findings of this paper have an impact on: (i) Cyclist risk awareness and route decisions. (ii) Policy-making for the improvement of transport infrastructure and encouragement of environmentally friendly transport means such as bikes by new cyclists and tourists.

Future work includes the expansion of the risk and discomfort estimation with exposure and vibration measures [4,17], the application of the introduced process to other cities than Zürich, city comparisons [9], the influence of other traffic in cycling safety as well as the design of traffic simulation models for the participatory multi-objective optimization of traffic flows [1,12,16]. An additional interesting avenue for research and applicability of the software artifact is other accident types than self-caused accidents, for instance the ones caused by L/R turns.

References

1. Amini S, Gerostathopoulos I, Prehofer C (2017) Big data analytics architecture for real-time traffic control. In: 2017 5th IEEE international conference on models and technologies for intelligent transportation systems (MT-ITS), pp 710–715. IEEE
2. Behrendt F (2016) Why cycling matters for smart cities. Internet of bicycles for intelligent transport. J Transp Geogr 56:157–164
3. Bíl M, Andrášik R, Janoška Z (2013) Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation. Accid Anal Prev 55:265–273
4. Bíl M, Andrášik R, Kubeček J (2015) How comfortable are your cycling tracks? A new method for objective bicycle vibration measurement. Transp Res Part C Emerg Technol 56:415–425
5. Box G, Cox D (1964) An analysis of transformations. J R Stat Soc Ser B Methodol 26(2):211–252
6. Chen YC (2017) A tutorial on kernel density estimation and recent advances. Biostat Epidemiol 1(1):161–187
7. Hübner K, Schünemann B, Schilling T, Radusch I (2017) On assessing road safety aspects of a cycling router application. In: 2017 15th International conference on ITS telecommunications (ITST), pp 1–8. https://doi.org/10.1109/ITST.2017.7972199
8. Kahle D, Wickham H (2013) ggmap: spatial visualization with ggplot2. R J 5(1):144–162
9. Kaygisiz Ö, Hauger G (2017) Network-based point pattern analysis of bicycle accidents to improve cyclist safety. Transp Res Rec 2659(1):106–116

10. Molino JA, Kennedy JF, Johnson PL, Beuse PA, Emo AK, Do A (2009) Pedestrian and bicyclist exposure to risk: methodology for estimation in an urban environment. Transp Res Rec 2140(1):145–156. https://doi.org/10.3141/2140-16

11. World Health Organization (2015) Global status report on road safety 2015. World Health Organization, Geneva

12. Pournaras E, Pilgerstorfer P, Asikis T (2018) Decentralized collective learning for self-managed sharing economies. ACM Trans Auton Adapt Syst 13(2):10

13. Rivara F, Thompson D, Thompson R (1997) Epidemiology of bicycle injuries and risk factors for serious injury. Inj Prev 3(2):110–114

14. Rodgers GB (1995) Bicyclist deaths and fatality risk patterns. Accid Anal Prev 27(2):215–223

15. Salahub C, Pournaras E, Castells Graells D (2019) Bicycle route comforts and risks. https://doi.org/10.6084/m9.figshare.8118311.v1. https://figshare.com/articles/Bicycle_Route_Comforts_and_Risks/8118311/1

16. Schmid S, Gerostathopoulos I, Prehofer C, Bures T (2017) Self-adaptation based on big data analytics: a model problem and tool. In: 2017 IEEE/ACM 12th international symposium on software engineering for adaptive and self-managing systems (SEAMS), pp 102–108. IEEE

17. Turner S, Sener IN, Martin ME, Das S, Hampshire RC, Fitzpatrick K, Molnar LJ, Colety M, Robinson S, Shipp E, Wijesundera RK (2017) Synthesis of methods for estimating pedestrian and bicyclist exposure to risk at areawide levels and on specific transportation facilities. Technical report, United States. Federal Highway Administration. Office of Safety. https://rosap.ntl.bts.gov/view/dot/36098

18. Wickham H (2016) ggplot2: elegant graphics for data analysis. Springer, Berlin