



# A data-driven approach for assessing biking safety in cities

Sara Daraei<sup>1\*</sup>, Konstantinos Pelechrinis<sup>1</sup> and Daniele Quercia<sup>2</sup>

\*Correspondence:

[daraei.sara@pitt.edu](mailto:daraei.sara@pitt.edu)

<sup>1</sup>School of Computing and Information Science, University of Pittsburgh, Pittsburgh, PA, USA  
Full list of author information is available at the end of the article

## Abstract

With the focus that cities around the world have put on sustainable transportation during the past few years, biking has become one of the foci for local governments globally. Cities all over the world invest in biking infrastructure, including bike lanes, bike parking racks, shared (dockless) bike systems etc. However, one of the critical factors in converting city-dwellers to (regular) bike users/commuters is safety. In this work, we utilize bike accident data from different cities to model the biking safety based on street-level (geographical and infrastructural) features. Our evaluations indicate that our model provides well-calibrated probabilities that accurately capture the risk of a biking accident. We further perform cross-city comparisons in order to explore whether there are universal features that relate to cycling safety. Finally, we discuss and showcase how our model can be utilized to explore “what-if” scenarios and facilitate policy decision making.

**Keywords:** Urban informatics; Bike safety; Modeling

## 1 Introduction

Transport engineers and urban planners have come to realize during the past few years that if we are to realize the vision of *smart cities*, that is, cities that are livable, sustainable and resilient, we need to move away from the car-centric mobility and turn to more sustainable modes of transportation. As a result cities have indeed turned to alternative modes of transportation, while promoting multi-modal mobility. Bicycles offer a promising transportation alternative to private vehicles, especially in areas with congestion, poor air quality, and high fuel prices [1]. Therefore, one of the major emphasis has been placed on urban biking as an alternative mode of transportation [2]. As a result during the past years the relevant infrastructure (e.g., bike lanes, bike parking racks, etc.) and services (e.g., shared bike systems) have enjoyed a significant growth [3]. This investment has resulted in an increased (perceived and/or actual) safety, either directly or through network effects [4], which consequently has a positive feedback on the propensity of people to ride a bicycle for their commute [5–9].

However, despite this increase in ridership the fraction of commuters that regularly use bikes is still relatively small as compared to other modes of transportation [10]. It is thus, important to understand what factors are related with bike safety and how we can make

© The Author(s) 2021. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

riding safer in a city. While there exists literature that aims at understanding through surveys how people perceive biking safety,<sup>1</sup> in this study we take a data-driven approach, using bike accident data from the cities of London, Boston and Pittsburgh. We further collect street-level features from OpenStreetMap that capture both the geography as well as, the bike infrastructure on the streets of these cities. Using these data we build a model that quantifies biking safety as the relative probability of a severe biking accident in a given location. Our results indicate that the probability output of our models is well-calibrated both when evaluated (i) on a test set from the same city that the model was trained on, as well as, when evaluated (ii) on a dataset from a different city. The latter, i.e., a cross-city study/evaluation, is a piece that is missing in many studies in the area of urban computing and analytics despite its importance. If we are to really understand various urban phenomena and their *universal* patterns (if any) across cities, similar cross-city analysis is necessary.

In brief, our results indicate that the presence of (protected) bike lanes is associated with a significant improvement in biking safety—as one might have expected. Furthermore, the speed limit, street topology (straight/curved), and the distance from an intersection are good predictors for the severity of a biking accident as well. While these results might not be surprising, and in fact they might even be expected, our work quantifies these relationships. This further enables their use in “what-if” analysis that can facilitate local government decision making as we elaborate on later. In summary, the main contributions of our work are:

- We develop a bike safety model using public bike accident and map datasets from 3 diverse cities.
- We perform a cross-city evaluation in order to identify possibly universal patterns related with biking safety.
- We also showcase how these models can be used by policy makers to evaluate the current infrastructure and decide on updates that maximize bike safety (possibly under other constraints).

The rest of the paper is structured as follows: Sect. 2 discusses related to our study literature and further differentiates our work. Section 3 describes the datasets used in our analysis as well as the biking safety model developed. We further provide our model evaluations in the same section. Section 4 shows how our model can be used for facilitating decisions on infrastructure updates that maximize biking safety under various constraints. Finally, Sect. 5 discusses the limitations of our study, while Sect. 6 concludes our work.

## 2 Related studies

In this section we will discuss relevant to our study research and position our study within this literature.

For the past few decades, transportation planners have been trying to identify how an urban environment can attract more bike commuters. For example, Clarke [11] explores the key factors in having a bike-friendly city. While Clarke focuses on the institutionalization of bicycle programs within agencies, safety is identified as an important factor that agencies will have to consider and overcome in order for biking to be fully integrated to the rest of the urban transportation network. As a result there are several studies in the literature

---

<sup>1</sup>We discuss this and other relevant literature extensively in the following section.

that aim at assessing the safety (or risk) of biking in the city and identifying characteristics of the street network that could help improve safety. A main differentiating point between these studies is the way the concept of safety is quantified. There are three ways that have been used to examine the concept of safety:

- *Actual safety*: In this case the risk of biking is directly assessed through data that capture both accidents/fatalities as well as exposure of bicyclists to these risks (e.g., bike trip counts).
- *Perceived safety*: In this case risk of biking is assessed through user surveys, and hence, provide a subjective view of the road users.
- *Inferred safety*: When there is not a direct measure of quantifying the risk of biking, the latter can be inferred through indirect measures, such as the relative position or speed of motor vehicles to bicycles.

In order to calculate the actual safety, studies have relied on exposure estimates from national surveys (e.g., [12]) or through video taping of traffic (e.g., [13, 14]). An interesting approach for dealing with the problem of estimating the exposure volume (e.g., the number of trips a bicyclist has taken, the number of bike trips over a specific segment etc.) has been presented in [15, 16]. In these studies, the authors use a case-crossover approach, where data from bike accident locations are *matched* with a random point on the trajectory the bicyclist followed at the trip that led to the accident. The dataset that is created this way allows to examine the influence of infrastructure on injury risk, while ensuring strict control for external covariates (i.e., exposure to risk, cyclist traffic volume) and for personal and trip characteristics (e.g., propensity for risk-taking, time of day). However, while this case-crossover approach does not require knowledge of the exposure volume, it does require the full trajectories of bike trips, which can also be difficult to access.

Given the possible lack of detailed data on bicycle accident or risk exposure, especially during the beginning of biking expansion in cities,<sup>2</sup> researchers have also been interested in assessing and quantifying the perceived safety of biking. Even though actual and perceived safety can differ, understanding how people perceive biking safety is particularly important; it can provide simple ways with which transportation and city authorities can *nudge* more people to use bicycle for their commute. To that end several studies have performed user surveys to understand what street features commuters associate with safety and/or which street aspects cyclist pay attention to (e.g., [18–22]).

This research has led to the development of various indices that aim at quantifying the risk/safety associated with biking over a specific street segment. In the late 1980s, early 1990s, the efforts were mainly focused on utilizing a combination of perceived and inferred safety. Davis [23] was the first to develop a score (Bicycle Safety Evaluation Index) with the objective of estimating the probability of an accident over a specific segment. The index combines information for the number of lanes, pavement, speed limit and similar variables. However, the weighting of each of these variables is *subjective* and it does not incorporate bicycle trips volume as an exposure variable. Despite its shortcomings the Davis work was the basis of the roadway condition index (RCI) and the Florida's Bicycle Coordinator segment condition index (SCI) [24]. Landis [25] built on these theoretical models to develop the interaction hazard score (IHS), whose results are evaluated through surveys with bicyclists, essentially attempting to recreate the perceived safety.

---

<sup>2</sup>In fact, even today that cities provide open information/data on bicycle accidents, these are constraint only to the accidents that are reported, which can be as low as 14% [17].

**Table 1** Basic information for our bike accident data

City	Years covered	Sample size
London	2005–2017	3765
Boston	2010–2013	2157
Pittsburgh	2004–2017	1724

Many of these very early attempts to develop models for bike safety in a city had to deal with the absence of access to detailed data on bike accidents (and street level mapping) that we have today. Closest to our study is the work by Allen-Munley *et al.* [26], who used crash data from Jersey City to model the severity of a bicycle crash. In particular, the authors—similar to our study—build a logistic regression model for the severity of an accident using as independent variables various street features (e.g., speed limit, number of lanes etc.). The authors are focused more on a descriptive model, and hence, they do not evaluate the predictive power of their models. The latter is the main focus of our study, since it can drive educated recommendations for alterations in the infrastructure that potentially can improve biking safety. Furthermore, we also cross evaluate the models in different cities, in order to examine the presence of any *universal* patterns in terms of biking safety.

### 3 Data and methodology

In this section we will describe the datasets we used for our analysis as well as the biking safety model we developed.

*Bike Accident Data:* To perform our study we obtained data on traffic accidents that involved bicycles from three metropolitan areas across Europe and North America. While these data are obtained from different sources (i.e., the Open Data portal of each city) they all provide the information we need in order to build the safety model. In particular, we collected data from:

- *London:* London's bike accident dataset is obtained from CycleStreet<sup>3</sup> and covers the period between 2005 and 2017.
- *Boston:* Boston's dataset is obtained through Cambridge Open Data<sup>4</sup> and covers the period between 2010 and 2013.
- *Pittsburgh:* Pittsburgh's dataset is obtained through the Western Pennsylvania Regional Data Center<sup>5</sup> and covers the period between 2004 and 2017.

Table 1 provides some additional basic information from the data. While these datasets are compiled, curated and distributed by different entities, the subset of the information that we need for building a biking safety model is included in all three datasets. In particular, the tuple of interest is:  $\langle \text{Lat}, \text{Lon}, \text{Accident Severity} \rangle$ . While latitude and longitude is unambiguous information, the severity of an accident can be defined differently across the different cities. For example, in the London dataset the severity levels are specified as “slight”, “serious” and “fatal” while in the Pittsburgh dataset they are categorized as “not injuries”, “minor injury”, “moderate injury”, “major injury”, and “killed”. Finally, in the Boston dataset the severity is represented by a Likert-like scale between 0 (no injury) to 4 (fatal). In order to be able to build models that are comparable between the

<sup>3</sup><https://bikedata.cyclestreets.net>

<sup>4</sup><https://data.cambridgema.gov/Public-Safety/Crash-Map-2010-2013/4tke-xhvq>

<sup>5</sup><https://data.wprdc.org>

**Table 2** Label matching between datasets

Dataset	Existing labels	Unified labels
London	slight	slight
	severe	severe
	fatal	
Boston	0 (no injury)	slight
	1 (light injury)	
	2 (fair injury)	severe
	3 (serious injury)	
Pittsburgh	4 (fatal)	
	not injures	slight
	minor injury	
	moderate injury	severe
	major injury	
	Killed	

different cities, we have unified the labels for the severity of the accidents. In particular, we have used two labels, namely, (i) severe, and (ii) slight, and Table 2 shows the mapping between the labels in each city. We will further discuss the potential limitations from this in Sect. 5.

*Street Features Data:* Having the latitude and longitude of a traffic accident allows us to obtain various information about the location that could be correlated with the severity of an accident involving a bicycle. Using the OpenStreetMap's API we extract street-level features for each of the accident locations. These features include speed limit, the existence of a bikelane or not, the width and the length of the street segment, the hilliness, the topology of the street segment (i.e., whether it is straight or curved), as well as, the distance of the accident location from an intersection.

*Street Network:* An important factor that we would like to examine as part of our model is the traffic volume on each street segment. However, we do not have access to this information, and hence, we take an indirect approach using as proxy appropriate network features. In particular, we extract from OpenStreetMaps the actual street network of each city,  $\mathcal{N} = \{\mathcal{V}, \mathcal{E}\}$ . The set of nodes  $\mathcal{V}$  represents street intersections, while the set of edges  $\mathcal{E}$  represents street segments connecting these intersections, i.e., edge  $e_{ij} \in \mathcal{E}$  exists if there is a street segment between intersection  $i \in \mathcal{V}$  and intersection  $j \in \mathcal{V}$ . Using this network we calculate the edge betweenness centrality for every edge  $e \in \mathcal{E}$ . The edge betweenness  $\beta_e$  of edge  $e$  is simply the sum over all the possible pairs of nodes  $(i, j)$ , of the fraction of all pairs shortest paths between  $i$  and  $j$  that pass through edge  $e$ :

$$\beta_e = \sum_{i,j \in \mathcal{V}} \frac{\sigma_{ij}(e)}{\sigma_{ij}}, \quad (1)$$

where  $\sigma_{ij}$  is the number of all shortest paths between nodes  $i$  and  $j$ , while  $\sigma_{ij}(e)$  is the number of those that pass through  $e$ . The betweenness centrality of an edge  $e$  is essentially proportional to the probability that  $e$  will be part of one of the shortest paths between a randomly selected pair of nodes  $i$  and  $j$ . Hence, one can make a very plausible hypothesis that the betweenness centrality of a street segment in our network  $\mathcal{N}$  is a good proxy for the overall traffic over the corresponding segment.

### 3.1 Bike safety models

To reiterate, our goal is to build a model that will capture the safety  $s$  (or risk  $\rho$  depending on how it is viewed) of biking in an urban environment. Ideally, given data from bike accidents on a street segment a risk score would be assigned to the segment based on the number of these accidents. However, this number should be normalized with the total number of bike trips that go through this street segment. If we only know that there were 10 bike accidents over a street segment, this does not inform us about the biking risk  $\rho$  associated with the segment. It can be anywhere from extremely low (e.g., if there have been 1 million bike trips over the segment), to completely risky (e.g., if there have been only 10 bike trips in total over the segment). We could then use this normalized accident count as our dependent variable for a safety model (e.g., through a beta regression since the dependent variable would be bounded by 0 and 1). Nevertheless, traffic data are not easily available/accessible, and this problem is particularly pronounced for bike traffic. Furthermore, in this setting the choice of spatial aggregation could become challenging. While street segment might seem an appropriate aggregation, different street segments can have vastly different lengths, which can impact both the estimation of the independent variables themselves, as well as, the dependent variable. The latter could be dealt with, by normalizing with the length of the street, but then interpretation and model choice (since the dependent variable is not a probability anymore and beta regression would not be appropriate) would be challenging. Therefore, we rely on a slightly different modeling setting and definition of risk  $\rho$  that has also been used in existing literature [26].

In particular, we focus on a specific location  $l$  (that is, an actual latitude/longitude pair, rather than street segment or other spatial aggregation), and define the risk associated with this location  $\rho_l$  as the probability of an accident that happens at  $l$  being *severe*. We can think of this definition as a second order safety score. Simply put  $\rho_l$  is the conditional probability of an accident at  $l$  being severe given that an accident at  $l$  happened ( $A_l$ ), i.e.,  $\rho_l = \Pr[\Sigma_l | A_l, \mathbf{x}]$ . The aforementioned probability is also conditioned on the vector  $\mathbf{x}$ , which are the independent variables of our model as described in what follows. We model this probability through a logistic regression model, i.e.:

$$\Pr[\Sigma_l | A_l, \mathbf{x}] = \frac{1}{1 + e^{-\mathbf{a}^T \cdot \mathbf{x}}} \quad (2)$$

In this setting, every bike accident recorded is a data point associated with a binary variable (1: severe, 0: non-severe accident). Then for every point of bike accident we obtain the following features as our independent variables:

1. *Speed-limit* ( $v$ ): This variable captures the allowed maximum speed in the corresponding accident location.
2. *Street width* ( $w$ ): This features represents the width of street of corresponding accident location which also could be considered as an indicator for the number of lanes.
3. *Distance to closest intersection* ( $d$ ): This feature captures the distance to the closest intersection for the corresponding accident location.
4. *Hilliness level* ( $h$ ): This is a binary variable that represents the steepness of the corresponding location and can be either “hilly” or “flat”.
5. *Street topology* ( $\theta$ ): This binary variable demonstrates if the accident happened in a curved or straight street.

**Table 3** Logistic regression coefficients for the risk model built with the datasets from different cities

Variable	London	Boston	Pittsburgh
(Intercept)	4.372***	4.840***	2.979***
speed-limit	11.071***	6.228***	12.332***
Street width	8.937**	3.827*	6.382***
betweenness	5.093***	4.827***	3.082**
dist-intersect	-6.978***	-4.280**	-3.829**
(hilliness) hilly	7.832***	7.374**	10.736***
(topology) curved	12.932***	6.482***	3.278***
(bikelane) with	-14.891***	-18.349***	-11.272***
<i>Interaction terms</i>			
speed-limit. betweenness	11.742**	4.729**	3.952*
speed-limit. (bikelane) with	-21.922***	-23.749***	-23.739***
speed-limit. dist-intersect	4.281	4.839	3.398*

\*\*\* $p < 0.01$ , \*\* $p < 0.01$ , \* $p < 0.05$ .

6. *Bikelane* ( $b$ ): This binary variable captures the existence of bike-lane at the location of the corresponding accident.
7. *Street Betweenness* ( $\beta$ ): This feature captures the edge betweenness centrality of the street segment where the accident occurred in the street network. This variable can be thought of as a proxy for the traffic volume at the location of the accident.

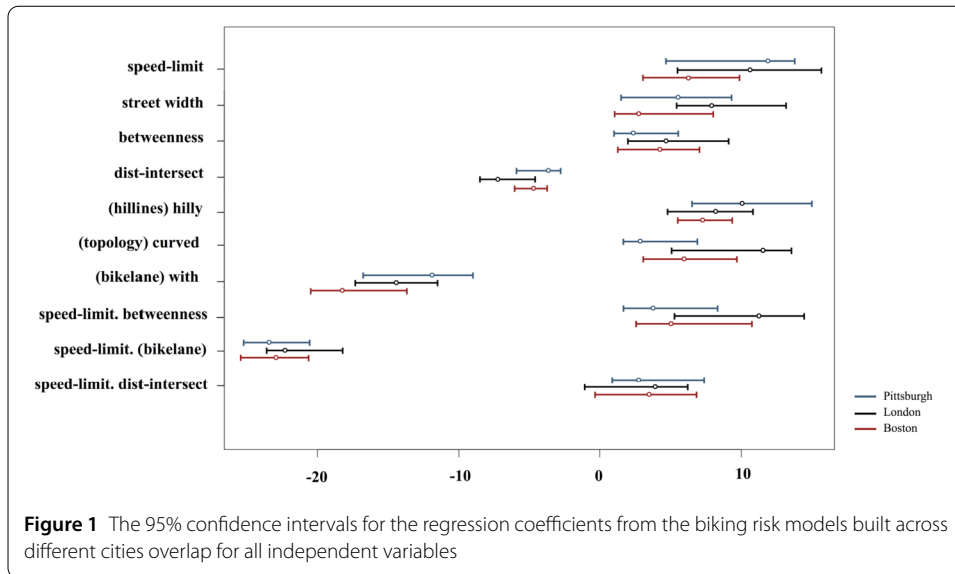
Table 3 presents the results obtained from each city. We train these models using the last 4 years that each dataset covers. The reason for restricting building the model with more (but older) data is twofold: (i) the street features 15 years ago are most probably (very) different than what they are now, and (ii) as people get more used to biking around the city and use improved biking equipment that exists today the risks associated can also change. Therefore, we want to restrict our model to only (fairly) recent data. As we can see all of the features are strongly correlated with the probability of a severe accident and the direction of these correlation is consistent across cities. For example, an accident that happens on a central street (i.e., large betweenness) is expected to have higher risk for severity, while the presence of bike lanes reduces this risk.

However, apart from the direction of the correlations, we are also interested in how different the various coefficients of the models are. Hence, in Fig. 1 we present the 95% confidence interval of the coefficients for each of the models. As we can see, even though the actual coefficients are different, their confidence intervals overlap, which essentially means that there is no statistical difference between the model coefficients.<sup>6</sup> In other words, biking safety appears to be correlated with street features in similar ways across the different cities we examined.

Finally, we selected a 1000 random locations from a different city (San Francisco, California) and applied each one of the risk models learnt to estimate the conditional probability of a severe bike accident on that location. Our goal with this experiment is to compare the differences in the predictions provided by the various models for the same points. Figure 2 presents the scatter plots of these predictions, where the two axes presents the risk probability based on the models from two different cities. As we can see the points follow closely the  $y = x$  line, which means that all models give similar risk probabilities for the same locations.

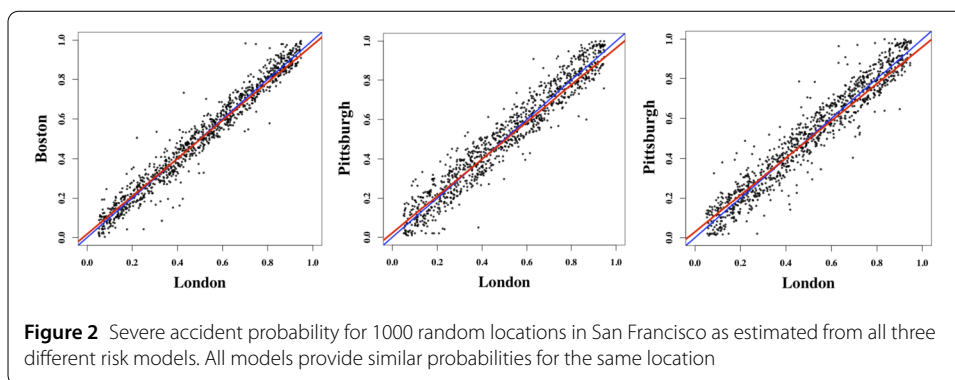
<sup>6</sup>We also present in Table 4 the  $p$ -values for the difference of the model coefficients.





**Table 4** Difference in regression coefficients and corresponding *p*-value

Variable	London vs. Boston	London vs. Pittsburgh	Boston vs. Pittsburgh
(Intercept)	-0.468 (0.471)	1.393 (0.373)	1.861 (0.251)
speed-limit	4.843 (0.103)	-1.261 (0.284)	-6.101 (0.061)
Street width	5.11 (0.126)	2.555 (0.275)	-2.555 (0.101)
betweenness	0.266 (0.201)	2.011 (0.052)	1.745 (0.071)
dist-intersect	-2.698 (0.275)	-3.149 (0.307)	-0.451 (0.065)
(hilliness) hilly	0.458 (0.347)	-2.904 (0.141)	-3.362 (0.094)
(topology) curved	6.451 (0.171)	9.654 (0.234)	3.542 (0.053)
(bikelane) with	3.458 (0.289)	3.619 (0.081)	-7.077 (0.242)
<i>Interaction terms</i>			
speed-limit. betweenness	7.013 (0.101)	7.791 (0.065)	0.777 (0.053)
speed-limit. (bikelane) with	1.827 (0.242)	-0.01 (0.218)	-0.23 (0.818)
speed-limit. dist-intersect	-0.558 (0.051)	0.883 (0.053)	1.441 (0.068)



### 3.2 Evaluating the models’ predictive power

Apart from the descriptive nature of the biking risk models explored above, we are also interested in their predictive power. To evaluate the predictive power of each model we split each dataset to a training and a testing set. In particular, for each city we use the first two years in our dataset for training and the last two as our out-of-sample testing set. Table 5 present the out-of-sample accuracy for each model, as well as, the Brier score. The latter



**Table 5** Out-of-sample accuracy and Brier score for the different models

City	Accuracy	BS	BS-reference	BSS
London	0.885	0.119	0.14	0.15
Boston	0.832	0.169	0.19	0.11
Pittsburgh	0.821	0.178	0.21	0.15

is a measure of probability calibration. Specifically, for a probabilistic model its classification accuracy paints only part of the picture. For example, two models  $M_1$  and  $M_2$  that both predict an accident will be severe—i.e.,  $\Pr[\Sigma_l|A_l, \mathbf{x}] > 0.5$ —they will have the same accuracy. However, if  $\Pr_{M_1}[\Sigma_l|A_l, \mathbf{x}] = 0.9$  and  $\Pr_{M_2}[\Sigma_l|A_l, \mathbf{x}] = 0.55$ , the two models have different probability calibration. Brier score [27] quantifies this calibration for each model. In particular, for the case of binary probabilistic prediction, the Brier score is calculated as:

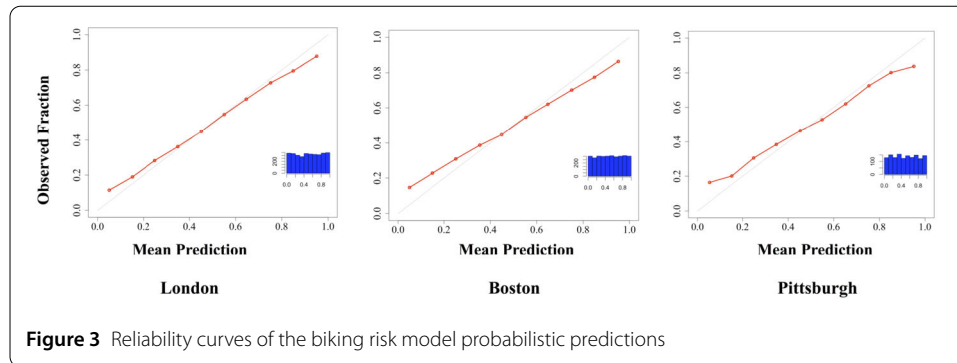
$$BS = \frac{1}{N} \sum_{i=1}^N (\pi_i - y_i)^2, \quad (3)$$

where  $N$  is the number of observations,  $\pi_i$  is the probability assigned to instance  $i$  of being equal to 1 and  $y_i$  is the actual (binary) value of instance  $i$ . The Brier score takes values between 0 and 1 and as alluded to above evaluates the calibration of these probabilities, that is, the level of confidence they provide. The lower the value of BS the better calibrated the output probabilities are. Continuing on the example above a 0.9 probability is better calibrated compared to a 0.55 probability (when the ground truth is label 1) and hence, even though  $M_1$  and  $M_2$  have the same accuracy,  $M_1$  is better calibrated (lower Brier score –0.01 compared to 0.2025). As we can see all models exhibit out-of-sample accuracy higher than 82%, while their calibration is very good as captured by the Brier score. In particular, the Brier score for each model is much lower as compared to the corresponding score for a reference model. The latter is a typical baseline Brier score used for evaluating the quality of a model, and is obtained through a baseline model that assigns to each data point the base probability of the positive class (in our case “severe” accident). Using the Brier score of the reference model  $BS_{ref}$ , we can further calculate the *skill* of probability estimates using the Brier Skill Score (*BSS*) Specifically, if  $BS_{ref}$  is the Brier score of the reference model, then *BSS* is calculated as:

$$BSS = 1 - \frac{BS}{BS_{ref}} \quad (4)$$

*BSS* will be equal to 1 for a model with perfect calibration, i.e.,  $BS = 0$ . A model with no *skill* over the reference model, will have a value of 0 since  $BS = BS_{ref}$ . If  $BSS < 0$ , then the model exhibits less skill than even the reference model.

Finally, we evaluate the accuracy of the probability output of each model by deriving the probability calibration curves for the test set in each case. In order to compute the accuracy of the predicted probabilities we would ideally want to have several bike accidents (e.g., 100) happen at the same location  $l$ . If the model assigned a 75% probability of an accident at location  $l$  being severe, then we would expect about 75 of the accidents observed in  $l$  to be severe. However, this is clearly not realistic to have (at least for the vast majority of the locations) and hence, in order to evaluate the accuracy of the probabilities we will use

**Table 6** Accuracy and Brier score for out-of-city evaluations

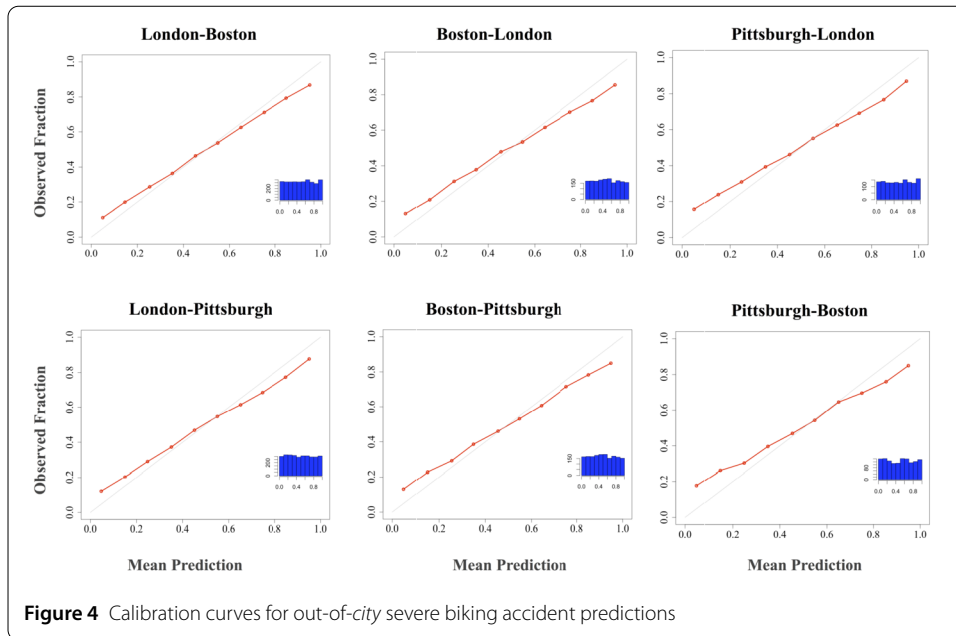
Training city	Testing city	Accuracy	BS	BS-reference	BSS
London	Boston	0.857	0.142	0.17	0.16
London	Pittsburgh	0.822	0.177	0.21	0.16
Boston	London	0.806	0.193	0.22	0.12
Boston	Pittsburgh	0.797	0.222	0.24	0.08
Pittsburgh	London	0.786	0.210	0.24	0.13
Pittsburgh	Boston	0.763	0.236	0.25	0.06

all the accidents in our dataset. In particular, if the predicted probabilities were accurate, when considering all the accidents  $\mathcal{A}$  where a severe incident was predicted with a probability of  $x\%$ , then a severe accident should have been observed in (approximately)  $x\%$  of  $\mathcal{A}$ . Given the continuous nature of the probabilities we quantize them into groups that cover a 10% probability range. Figure 3 presents the predicted probability of a severe accident on the x-axis, while the y-axis presents how many of these accidents were indeed severe. Furthermore, we present in the inset figure for each calibration curve the distribution of the predicted probabilities, which as we observe cover fairly uniformly the whole range of probabilities. As we can see the calibration curve is very close to the  $y = x$  line, which practically means that the predicted probabilities capture fairly well the actual biking risk probabilities.

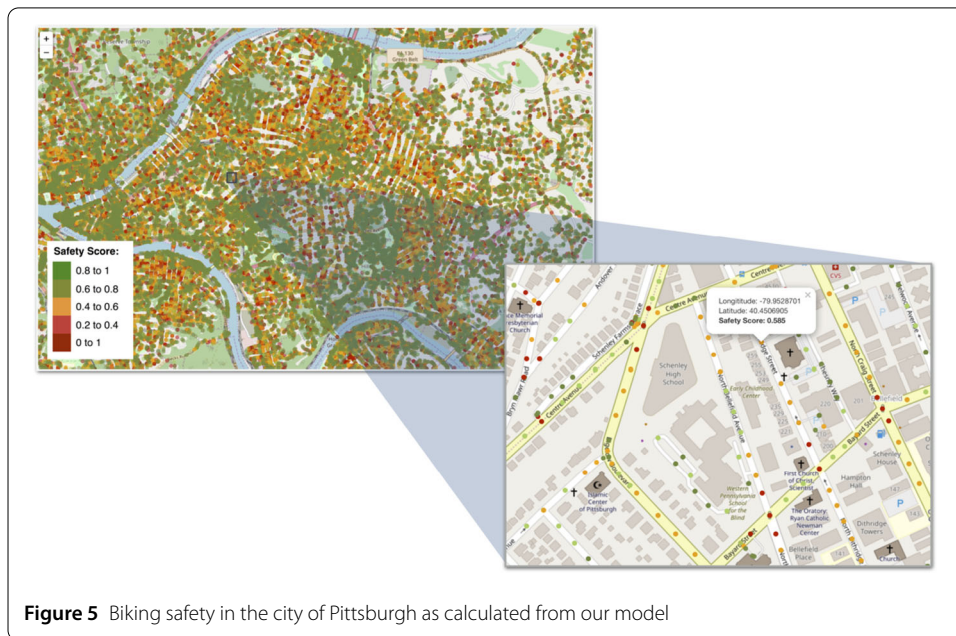
### 3.3 Cross-city models

In Sect. 3.1, we compared the coefficients for the safety models we built for the different cities. Here we want to evaluate the ability of a model trained on a specific city to predict the severity of accidents in a different city. These cross-city evaluations and comparisons are largely absent from the current literature in urban informatics, but are very important. The ability to predict the severity of accidents in city A based on a model that was trained with data from city B, is a good indicator that there might be universal patterns in biking safety. Simply put, knowledge obtained from a specific city might be transferable (and generalizable) to another city.

In particular, we use the model trained with data from each city in our study to predict the severity of accidents in the other two cities in our study (6 pairs in total). Table 6 presents the accuracy and the Brier score for these experiments, while Fig. 4 shows the reliability curves for these predictions. As we can see, the accuracy, while reduced as compared to the out-of-sample performance for the same city, is still high. The same is true for the Brier score (and the corresponding Brier skill scores). In particular, the performance



**Figure 4** Calibration curves for out-of-city severe biking accident predictions



**Figure 5** Biking safety in the city of Pittsburgh as calculated from our model

of a model trained on London data and evaluated on Boston data—and vice versa—exhibit almost as good performance as the within city evaluations. The dataset from the city of Pittsburgh is much smaller and this could impact the model quality, especially when used in this cross-city experiment. Finally, Fig. 4 presents the calibration curves for the different pairs of training-test cities. As we can see again, the calibration of the probabilities are good and the curve is close to the  $y = x$  line.

#### 4 Application demonstration: improving bike safety under constraints

This type of models allow us to evaluate the biking safety within a city and can facilitate policy decisions with respect to infrastructure updates. For example, Fig. 5 presents the

interactive map of results we obtained from our model for the whole city of Pittsburgh. By clicking on each point, users are able to obtain its safety score. In particular, for every point on the street network we estimate the safety  $s$  (which is  $1 - \rho$ ) associated with this point and visualize it on the map. Safe locations (i.e., low risk of severe accident) are colored green, while risky locations are colored red. Using these results one can start identifying locations that are not as safe as needed and explore what infrastructure changes (e.g., widening of a street, installation of a bike lane etc.) are going to have the maximum impact on biking safety, possibly under budgetary (and policy) constraints.

For example, Fig. 6(a) presents a three square miles area located within the city of Pittsburgh. The selected area does not currently include separate bikelanes for cyclists' transportation. The average safety score obtained for this area is  $s = 0.54$ .

Let us assume that we would like to enhance biking safety for this area and we have budget for adding bike lanes. Furthermore, we have the constraint that we cannot install bike lanes on local streets (which typically have low traffic anyway). Figure 6(b) presents the safety scores for the same area but now assuming that bike lanes have been installed in all of the non-local streets. Now the average safety score has increased to  $s = 0.68$ . This corresponds to an approximately 26% increase in the average biking safety within the area by *simply* adding bike lanes in non-local streets.

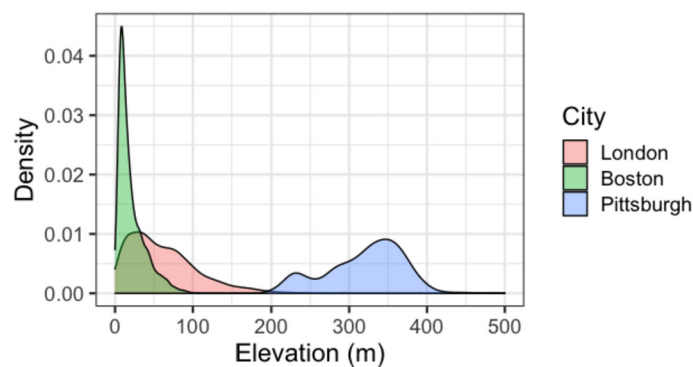
Obviously, with cities being cash-strapped today, they cannot just install bike lanes, or widen all the streets. However, models similar to the ones we developed in this study, allow policy makers to better understand the correlations between the characteristics of



the existing infrastructure and how one could potentially improve the safety for bikers. In fact, one could explore a variety of (viable) options and decide which one is the *optimal* based on the criteria of interest (e.g., trade off between safety and cost etc.).

## 5 Discussion and limitations

Our study contributes to the literature of urban informatics and more specifically to the area of computational transportation modeling. A large portion of the literature that lays at the intersection of transportation and cycling is focused on studying the commute behavior of cyclists, policies and infrastructure that can nudge people towards cycling, as well as, the benefits—both personal and societal—from increased levels of biking in a city (e.g., [28–36], with the list not being exhaustive). As mentioned earlier bike safety is related with increased levels of biking, and even though causality most probably runs in both directions [28, 37], understanding the association between the urban infrastructure and cycling safety is crucial if cities are to rely more on biking in the future. While models for biking safety have been developed in the past they are typically based on a small number of observations and most importantly cover a single city/area. In our study we aim at identifying aspects of biking safety that are similar across cities. While the three cities we examined are not representative of all—or even most of—the cities worldwide, they exhibit some very substantial differences in terms of both their biking infrastructure, as well as, their geography. For example, we used MapQuest’s elevation API and obtained a uniformly random sample of elevation for points on the street network of each city (approximately 15,000 points in each city). Figure 7 presents the distributions of the elevation for the three cities. What is particularly interesting in this case in terms of biking is the variability of the elevation across the street network. As we can see the variability in Boston is much lower as compared to London and Pittsburgh. The standard deviation of the elevation is approximately 57, 43 and 17 meters for Pittsburgh, London and Boston respectively. While elevation is only one aspect of a city’s geography, it is a particularly important one for biking and bike usage [38]. Furthermore, the bike infrastructure in these cities is different. Based on our OpenStreetMap data, the fraction of street length equipped with biking infrastructure (dedicated lanes and protected tracks) varies. In London, this fraction is 9.2%, which is in stark contrast to only 5.6% and 2.8% for Boston and Pittsburgh respectively. Of course, there are also deeper differences (at least between London and the two US cities in



**Figure 7** The distribution of elevation for a random sample of points on the street network for the three cities studied



our study) that relate to the overall structure and planning of the city, that is, the division between traditional cities and the radiant cities model (with the latter being the model for many American cities). While of course we cannot claim that our models capture globally universal patterns, the differences across the three cities studied provide support for their generalizability potential.

Furthermore, our street level data were collected from OpenStreetMap. While OpenStreetMap is expected to provide high quality information in big cities like London, Boston and Pittsburgh, it is still a crowdsourcing platform and hence, it is possible to still have incomplete/erroneous information. However, the general trends identified are not expected to be significantly impacted by the crowdsourcing nature of OpenStreetMap. One of the possible limitations of our study is that some of the street features we collected from OpenStreetMap are contemporary and they might have been different when the actual accident data were recorded. For example, the speed limit, or the presence of a bikelane might have changed over the years. We have tried to account for that by analyzing only recent data. Similarly, while we have used the street betweenness  $\beta$  as a proxy for the traffic on a street, this does not account for the temporal nature of the traffic. Simply put, even though betweenness might be a good proxy for the average traffic over a day, it is not able to capture its temporal patterns. Moreover, the original severity labels for an accident have different categories. Although we unified them as demonstrated in Table 2, there is still a subjective element on how the original data were classified.

Finally, other classification models (e.g., tree classifiers, RNNs etc.) could possibly perform better in predicting the severity of a bike accident. However, we wanted a model that is easily interpretable. This is particularly important for policy makers that want to make decisions based on the information provided by the model. At the same time more independent variables, e.g., features obtained from visual cues—presence of trees, head-in parking, presence of billboards etc.—could provide further information and further improve the quality of the model. We plan to explore similar features in future studies.

## 6 Conclusions

In this work our objective is to build an appropriate model that will allow us to understand the relationship between street infrastructure features and biking safety. To achieve our objective we obtain traffic accident data which involved a bicycle from three metropolitan areas, namely, London, Boston, and Pittsburgh. We further collected street features for the locations of the accidents and built a logistic regression model for each city for the probability of a severe accident, conditional to the presence of an accident. Our results indicate that these models are transferable between cities (at least between the ones we examined). In particular, the coefficients are statistically the same in all three models, while each model has very good predictive power for accidents that happened in a different city. Finally, we show how our models can be useful to policy makers by evaluating the current state of bike safety within a city and analyzing “what-if” scenarios for infrastructure updates.

### Acknowledgements

None.

### Funding

This project was partially funded by the PittSmartLiving project (NSF CNS-1739413).

### Abbreviations

BS, Brier Score; BSS, Brier Skill Score; IHS, Interaction Hazard Score; RCI, Roadway Condition Index; SCI, Segment Coordination Index.

### Availability of data and materials

Data used in this paper are available at: <https://github.com/daraei/BikeSafety/tree/master/BikeSafety>.

### Competing interests

The authors declare that they have no competing interests.

### Authors' contributions

SD run the analysis and wrote the paper. KP conceptualized the problem, run the analysis and wrote the paper. DQ conceptualized the problem and wrote the paper. All authors read and approved the final manuscript.

### Author details

<sup>1</sup>School of Computing and Information Science, University of Pittsburgh, Pittsburgh, PA, USA. <sup>2</sup>Nokia Bell Labs, Cambridge, UK.

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 14 January 2020 Accepted: 3 February 2021 Published online: 03 March 2021

### References

1. Klobucar M, Fricker J (2007) Network evaluation tool to improve real and perceived bicycle safety. *Transp Res Rec* 2031:25–33
2. Martens K (2004) The bicycle as a feeder mode: experiences from three European countries. *Transp Res, Part D, Transp Environ* 9(4):281–294
3. StreetsBlogUSA (2016) Report: As cities add bike lanes, more people bike and biking gets safer. <https://usa.streetsblog.org/2016/07/20/report-as-cities-add-bike-lanes-more-people-bike-and-biking-gets-safer/>. [Online; accessed 19-December-2019] (2016)
4. Jacobsen PL (2003) Safety in numbers: more walkers and bicyclists, safer walking and bicycling. *Injury Prev* 9(3):205–209
5. Baker L (2009) How to get more bicyclists on the road. *Sci Am* 301:28–29
6. The Guardian (2014) Cycling in cities: the search for the world's most bike-friendly metropolis. <https://www.theguardian.com/cities/bike-blog/2014/mar/04/cycling-cities-search-bike-friendly-metropolis>. [Online; accessed 19-December-2019]
7. CityLab (2014) Protected bike lanes aren't just safer, they can also increase cycling. <https://www.citylab.com/transportation/2014/06/protected-bike-lanes-arent-just-safer-they-can-also-increase-cycling/371958/>. [Online; accessed 19-December-2019]
8. Jacobsen PL, Racioppi F, Rutter H (2009) Who owns the roads? How motorised traffic discourages walking and bicycling. *Injury Prev* 15(6):369–373
9. Pucher J, Buehler R (2008) Making cycling irresistible: lessons from the Netherlands, Denmark and Germany. *Transp Res* 28(4):495–528
10. Witlox F (2004) Evaluating bicycle-car transport mode competitiveness in an urban environment. An activity-based approach Frank Witlox & Hans Tindemans. *Editorial Board* 10(4):32–42
11. Clarke A (1992) Bicycle-friendly cities: key ingredients for success. *Transp Res Rec* 1372:71
12. Rodgers GB (1995) Bicyclist deaths and fatality risk patterns. *Accid Anal Prev* 27(2):215–223
13. Harkey DL, Reinfurt DW, Knuiam M, Stewart JR, Sorton A (1998) Development of the bicycle compatibility index: a level of service concept, final report. Technical report, United States, Federal Highway Administration
14. Hunter W, Stewart J, Stutts J (1999) Study of bicycle lanes versus wide curb lanes. *Transp Res Rec* 1674:70–77
15. Teschke K, Harris MA, Reynolds CC, Winters M, Babul S, Chipman M, Cusimano MD, Brubacher JR, Hunte G, Friedman SM et al (2012) Route infrastructure and the risk of injuries to bicyclists: a case-crossover study. *Am J Publ Health* 102(12):2336–2343
16. Harris MA, Reynolds CC, Winters M, Crompton PA, Shen H, Chipman ML, Cusimano MD, Babul S, Brubacher JR, Friedman SM et al (2013) Comparing the effects of infrastructure on bicycling injury at intersections and non-intersections using a case-crossover design. *Injury Prev* 19(5):303–310
17. Gaarder P (2009) Bicycle accidents in Maine: an analysis
18. van der Waerden P, Borgers A, Timmermans H (2004) Cyclists' perception and evaluation of street characteristics. In: 83rd annual meeting of the transportation research board (A3B07-committee on bicycle transportation), Washington
19. Badgett SI, Niemeier DA, Rutherford GS (1993) Bicycle commuting deterrents and incentives: a survey of selected companies in the greater Seattle area. In: 72nd annual meeting of the transportation research board, Washington
20. Goldsmith SA (1992) Reasons why bicycling and walking are and are not being used more extensively as travel modes, vol. 1. Federal Highway Administration
21. Aultman-Hall L, Adams M Jr (1998) Sidewalk bicycling safety issues. *Transp Res Rec* 1636:71–76
22. Sharples R (1999) Use of main roads by utility cyclists in urban areas. *Traffic Eng Control* 40(1):18–22
23. Davis WJ (1987) Bicycle safety evaluation. Ph.D. thesis, Auburn University
24. Epperson B (1994) Evaluating suitability of roadways for bicycle use: toward a cycling level-of-service standard, vol. HS-042 007
25. Landis BW (1994) Bicycle interaction hazard score: a theoretical model. Technical report



26. Allen-Munley C, Daniel J, Dhar S (2004) Logistic model for rating urban bicycle route safety. *Transp Res Rec* 1878:107–115
27. Brier GW (1950) Verification of forecasts expressed in terms of probability. *Mon Weather Rev* 78(1):1–3
28. Buehler R, Pucher J (2012) Cycling to work in 90 large American cities: new evidence on the role of bike paths and lanes. *Transportation* 39(2):409–432
29. Pucher J, Dill J, Handy S (2010) Infrastructure, programs, and policies to increase bicycling: an international review. *Prev Med* 50:106–125
30. Macmillan A, Connor J, Witten K, Kearns R, Rees D, Woodward A (2014) The societal costs and benefits of commuter bicycling: simulating the effects of specific policies using system dynamics modeling. *Environ Health Perspect* 122(4):335–344
31. Handy S, Van Wee B, Kroesen M (2014) Promoting cycling for transport: research needs and challenges. *Transp Rev* 34(1):4–24
32. Buehler R, Dill J (2016) Bikeway networks: a review of effects on cycling. *Transp Rev* 36(1):9–27
33. Stehlin J (2015) Cycles of investment: bicycle infrastructure, gentrification, and the restructuring of the San Francisco bay area. *Environ Plan* 47(1):121–137
34. Brown BB, Smith KR, Hanson H, Fan JX, Kowaleski-Jones L, Zick CD (2013) Neighborhood design for walking and biking: physical activity and body mass index. *Am J Prev Med* 44(3):231–238
35. Pelechrinis K, Zacharias C, Kokkosis M, Lappas T (2017) Economic impact and policy implications from urban shared transportation: the case of Pittsburgh's shared bike system. *PLoS ONE* 12(8):0184092
36. Braun LM (2018) Geographies of (dis) advantage in walking and cycling: perspectives on equity and social justice in planning for active transportation in us cities. PhD thesis, University of North Carolina at Chapel Hill
37. Jacobsen PL (2015) Safety in numbers: more walkers and bicyclists, safer walking and bicycling. *Injury Prev* 21(4):271–275
38. Kim I, Pelechrinis K, Lee AJ (2020) The anatomy of the daily usage of bike sharing systems: elevation, distance and seasonality. In: *ACM SIGKDD workshops on urban computing*

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

---

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)

---