Predicting Perceived Cycling Safety Levels Using Open and Crowdsourced Data

Jiahui Wu, Lingzi Hong and Vanessa Frias-Martinez
University of Maryland
Email: {jeffwu, lzhong, vfrias}@umd.edu

***Abstract*—Cycling communities have been related to lower obesity rates and lower stress levels. Nevertheless, one of the main obstacles to increase ridership in cities is the lack of information regarding perceived cycling safety at the street level. City planners have typically used extensive road network and traffic information to approximate cycling safety levels. However, this approach requires the deployment of expensive sensors thus making it hard for many cities to get access to accurate cycling safety maps. In this paper, we evaluate several methods to predict urban cycling safety at the street level, exclusively using public information from open and crowdsourced datasets. We evaluate the proposed approach in the city of Washington D.C. and achieve F1 scores of 66%, 70% and 88% when five, four or three different cycling safety levels are considered.**

***Keywords*-cycling safety; spatio-temporal analysis; open data**

I. Introduction

The benefits of cycling are well studied: reduction in urban pollution [1]; savings in healthcare costs due to daily exercise [2]; or improvements in workforce accessibility for low-income communities [3], among others. As a result, cities have created bike lanes, started bike-shared systems and supported bike to work programs in the past years. However, as bicyclists take the streets, bicycle safety is increasingly becoming an important concern [4]. One of the main obstacles to decrease the number of bicycle crashes is the lack of information regarding perceived cycling safety at the street level. Building bicycle-friendly communities that promote an increase in ridership will require accurate, citywide cycling safety maps. Such maps would allow policy makers to make informed decisions on the optimal distribution of budgets across road improvements and would back bicycle activists in their arguments.

Local departments of transportation (DoTs) use numeric scales *e.g.,* from 1 to 5, to measure cycling safety at the street level [5]. Safety levels are typically computed using different estimation models that require extensive information including features such as daily traffic or average speed [6]. However, accessing such information entails setting up expensive sensors thus highly limiting cycling safety maps to cities with economic resources, and to only a handful of streets where such sensors are deployed. In an attempt to make cycling safety maps accessible to a larger number of cities and for a larger number of streets, we propose to use information exclusively extracted from open and crowdsourced datasets as proxies to predict the cycling safety levels of urban streets. We will work with two types of information that have been reported to play a role in cycling safety perception, social features and built-in environment characteristics [7]. Social features which characterize citizen behaviors that take place in streets *e.g., crime*, can be extracted from cities' open data repositories (like *NYC open data* or *London Datastore*); while built-in environment characteristics that describe street features such as the presence of cycling facilities, can be extracted from crowdsourced platforms like Open Street Map. In this paper, we will evaluate various methods to predict perceived cycling safety levels at the street level using social and built-in environment features from Washington, D.C.; and we will show that we can lower the bar in the access to comprehensive and accurate cycling safety maps for many cities worldwide only using (a) open data repositories, which are available for over 2600 cities worldwide [8], and (b) crowdsourced data from Open Street Map, which is available for over 4 million small- to mid-sized cities [9].

Cities interested in using the cycling safety predictive methods we propose, will also require access to cycling safety ground truth levels so as to train their own local models to achieve the highest prediction accuracies. For that purpose, we have designed an open-source, crowdsourced rating platform which uses cycling videos recorded by real cyclists to collect perceived safety ratings from cyclists with different levels of expertise. These ratings, together with the social and built-in environment features extracted from open and crowdsourced datasets, will be used to assess the accuracy of various machine learning approaches in predicting perceived cycling safety levels.

II. Related Work

Research in transportation planning has focused on the design of quantitative models to estimate cycling safety levels [6]. One of the first models was the bicycle safety index rating (BSIR) which related bicycle safety to the physical and operational features of the roadway, using variables such as annual daily traffic, speed limit, lane width or pavement conditions to estimate cycling safety levels [10]. The Bicyclist Stress Levels index (BSL) was introduced by Sorton and Walsh who determined that stress levels (characterized from one to five) suffered by cyclists could be measured as a function of peak-hour traffic volume, motor vehicle speeds and lane width [11]. Building on all these models, the Bicycle Level of Service (BLOS) was the first measure to introduce the presence of a stripe separating

978-1-5386-5035-6/18/$31.00 ©2018 IEEE

motor vehicle from bicycle lanes as an important factor in determining cycling safety [5]. From BSIR to RCI or BLOS, all these safety estimation models suffer from two important limitations. First, they require a large set of variables that cities might not have the ability to collect at all, or can collect but only for a limited number of streets e.g., traffic information requires the deployment of speed sensors that due to their cost can only be put in selected places. Second, many of these estimation models are evaluated using a manual approach *i.e.,* the accuracy of the safety levels output by the models is checked through explorations of the physical environment by cycling safety experts rather than through data-driven approaches that would require far less work. The approach we present in this paper overcomes these limitations by using features extracted from open and crowdsourced datasets that many cities are already collecting, thus lowering the bar in the access to cycling safety maps for many cities worldwide.

III. Cycling Safety Prediction

In this section, we present an approach that uses social and built-in environment features as proxies to predict perceived cycling safety at the street level. Since streets in cities can be long, we focus on the prediction of cycling safety at the street segment level, defined as the section of the road between two adjacent intersections. We frame the cycling safety prediction as a classification problem where given a set of n features $F_i = \{F(1), ..., F(n)\}$ that characterize a street segment i, and given a ground truth safety label L_i that characterizes the perceived cycling safety for that segment, we explore the accuracy of various classification methods M in the prediction of the perceived cycling safety exclusively using the available features across all segments in the geographical area under study *i.e.,* $\vec{L} = M(\vec{F})$.

A. Prediction Features

It has been shown that variables that characterize the built-in and social environment of a street play a role in the perceived cycling safety. For example, streets that are known to have a lot of on-street parking tend to be avoided mostly due to dooring concerns (the door of a parked car opens onto the cycling lane and hits a cyclist) [12]; or that roads in high crime areas are less favored by cyclists when choosing their cycling routes [13]. Informed by an extensive exploration of the related work in cycling safety and by the types of variables typically available in open and crowdsourced datasets, we have created a list of features as potential predictive proxies for street cycling safety levels in urban environments. We consider two types of features: social features and built-in environment features. Social features focus on the characterization of human behaviors that take place in the streets and that change over time *e.g.,* parking violations or crime rates, while built-in environment features provide an atemporal depiction of the network formed by all the streets in the city.

We consider the following social features (*a* to *e*): (a) **crime rates** have been found to play a role in cycling safety perception, with cyclists avoiding high crime areas unless there is no other route available [13]. Crime statistics per street segment, typically collected by police departments, are usually available in open data portals by total volume or by type of crime *e.g.,* liquor law violations, thefts or robberies. We will evaluate both representations as predictive proxies for the perceived cycling safety of a segment; (b) specific **points of interest** such as parks, residential building areas or malls have been identified as being connected to high crime rates [14]. In an attempt to incorporate these factors into the cycling safety prediction models, we will extract the points of interest (POIs) at a given street segment using Open Street Map; (c) **bicycle-related crashes** have been shown to play a role in safety perception with cyclists favoring roads that are known to have low numbers of cycling related crashes [15]. Crash statistics per street segment are typically available by total volume or by type of crash *e.g., collision with fixed car* or *hit and run*. We will explore the use of both to predict cycling safety levels at the street segment; (d) **311 requests** are typically available in open data portals. These are citizen-initiated requests to solve a specific problem, and they are collected by city halls through their 311 portals. We will only use 311 requests related to street conditions such as number of curb, light bulb or road bump repairs as proxies for road conditions, since these have shown to affect the perception of cycling safety [16]; and finally, (e) **parking volumes** have been shown to impact safety perception, with higher parking volumes associated to less safety [17]. Although parking volumes are not typically available in open data portals, parking and moving violations characterized by their type are *e.g., distracted driving using cell phone* or *parked car obstructing sidewalk or driveway*. Thus, we will explore whether the volumes of parking and moving violations might help in predicting the perceived cycling safety.

On the other hand, we explore the following built-in environment features (*f* to *h*) as potential predictive proxies for cycling safety at the segment level: (f) **road network** variables including type of road (street, avenue, etc.), number of lanes, directionality and slope, which have been reported to play a role in cycling safety perception [18]. These features are available in Open Street Map, except for the slope which can be computed using Google's API Elevation Service, retrieving the elevation of several points in each segment; (g) **graph-based features** of the street segments in terms of centrality measures that quantify the importance of the segment in the overall road network *i.e.,* whether it is a central segment that is typically cycled through to go between any two points in the city, or more of an outlier segment. Related literature has shown that network centrality measures play a role in promoting cycling activities which

in turn create a critical mass that enhances the perception of cycling safety [19]. Road network maps can be retrieved from either Open Street Map or open data portals (as GIS resources). Using the SNAP package over the road networks will allow to evaluate various centrality measures such as degree, betweenness or page rank, among others, considering the road network of the city both as an undirected and directed graph (taking into account the direction of the traffic flow) [20]. Additionally, we will evaluate both primal and dual road network approaches that consider either each segment as an edge and each intersection as a node, or vice versa [21], [22]; and finally, (h) **presence of cycling facilities** and their type *e.g.,* dedicated bike lane or lane shared with traffic. These features, which can be extracted from Open Street Map, have also been shown to play a role in cycling safety perception [7], [11].

B. Crowdsourced Rating Platform

We have created a crowdsourced platform that cities can use to collect ground truth data from cyclists with respect to perceived cycling safety at the street segment level. The objective of the platform is to collect ground truth labels to be able to train and evaluate the accuracy of the prediction methods proposed. When users access the platform (http://www.cyclingsafety.umd.edu), they are first asked to voluntarily rate their cycling experience level by choosing among the four following standard options in the cycling literature: fearless, confident, interested or reluctant [23]. After that, the user will be shown $20s$ cycling videos recorded by actual cyclists and after each video, she will be prompted to provide a cycling safety rating between 1 (*too dangerous*) and 5 (*very safe*) as previously described in cycling safety literature [23]. Users are also asked about their familiarity with the route shown in the video, which will be used in the evaluation as a feature that might play a role in cycling safety perception.

Although platform users rate the perceived cycling safety conditions of the videos, we need to collect safety labels per street segment since that is the granularity of the proposed prediction methods. The platform internally computes the cycling safety levels at the street segment as follows. The videos shown in the platform have been recorded by cyclists with a bike-mounted camera. The recordings contain not only the video footage but also the GPS traces associated to the cycling trip. We use such GPS information to retrieve the street segments associated to a given video. However, such process is not straight forward since GPS sensors have errors, and more so in urban environments where when surrounded by tall buildings the GPS might loose signal or record a location quite far away from the actual visited point. As a result, we retrieve the list of street segments cycled using Mapbox's Map Matching API, which snaps fuzzy, inaccurate GPS traces to actual segments in the road network [24]. Internally, Mapbox uses the map-matching algorithm by Newson and Krumm, based on Hidden Markov Models (HMM) that find the most likely street segment in the network that is represented by the collected GPS location [25]. Since a video might include only portions of a given street segment i, the platform maintains internal lists that associate each video j to a set of street segments, with the percentage of each street segment covered within the video ($c_{i,j}$). Street segment safety levels (labels) are computed by (i) averaging all available participant ratings r for that segment across videos, with the ratings weighted by their segment coverage in the video *i.e.,* $L_i = (\sum_{j=1}^{m}(\sum_{q=1}^{n} c_{i,j} r_q)/n)/m$, where n is the number of segment ratings and m is the number of videos since a segment might partially or fully appear on multiple videos; and (ii) assigning it to its closest integer value in the range $[1\text{-}5]$.

IV. Evaluation

In this section, we evaluate the proposed cycling safety prediction approach using open and crowdsourced data for the city of Washington D.C. We first present the set of social and built-in environment features $F_i = \{F(1), ..., F(n)\}$ extracted for each street segment using Washington's D.C. open data portal and Open Street Map. Next, we describe the ground truth data collection of perceived cycling safety ratings using the crowdsourced rating platform deployed for Washington D.C. in collaboration with Washington Area Bicyclist Association (WABA). Finally, video safety ratings are transformed into street segment labels $L_i \in [1\text{-}5]$ and put together with the features to evaluate the accuracy of various cycling safety classification methods. We also evaluate the impact that sparsity and class imbalance might have on the accuracy of the classification methods.

A. Feature Extraction

We represent each street segment i in Washington D.C. as a set of built-in environment BE_i and social features S_i *i.e.,* $F_i = \{BE_i, S_i\}$. Built-in environment features are extracted from D.C.'s Open Street Map, and are represented as a number characterizing each of the features described in section III-A ((f)-(h)). Specifically, we extract 63 built-in environment features including 11 road network variables, 39 graph-based and 13 cycling facilities' variables *i.e.,* ($|BE_i| = 63$). On the other hand, the social features are extracted from D.C's open data portal and from Open Street Map (OSM). We retrieve time-stamped, geolocated events for the following 6 social features: crime, crash, 311 and parking and moving violations datasets for the past three years; and all the POIs in D.C. from OSM. Each social feature is divided into the following types: 11 types for crime data, 11 types for crash data, 72 for 311 requests, 10 for POIs, 36 different types of parking violations and 8 types of moving violations. We explore two representations for each social feature per street segment (except POIs): monthly average across all types and monthly average per type. The

main objective is to evaluate whether a more granular representation of the social features including volumes per type of event, rather than total volumes, has an impact on the final accuracy of the perceived cycling safety predictions. The monthly average across all types is computed as a number representing the average of the monthly feature values across the three years of data. Monthly average per type, on the other hand, is computed as an x-element vector where each element contains the average of the monthly feature values for each type across all three years. For example, the feature *crashes* is classified into 11 different types including assault, burglary or crime with dangerous weapon. Its monthly average would be computed as a number representing the average of all monthly crimes for the past three years; while the monthly average per type would be computed as a 11-element vector with each element representing the average of monthly crimes for a specific type of crime over the past 36 months. Thus, the final size of the social features' vectors will be $|S_i| = 6$ for the monthly average across types and $|S_i| = 148$ for the monthly average per type.

Measuring feature sparsity as the percentage of segments that have zero values for a given feature, we observe that the monthly averages across all types have very little sparsity, with values ranging between 0% and 9%. However, the monthly averages per type have larger sparsity values ranging from an average of 0.8% for crash violations to 1.5% for different types of parking or moving violations, and up to values higher than 60% for certain types of 311 reports. A comparison between the two monthly average representations, together with different classification methods and feature selection techniques, will allow to disentangle whether the sparsity of the feature vectors when using the monthly averages per type might affect the predictive accuracy. To preserve the interpretability of our models no sparse learning approach is used, since our main objective is to predict cycling safety levels and provide decision makers with actionable insights behind such levels.

Finally, it is important to clarify that the effect of social features on cycling safety perception might take place not only at the $lant, long$ coordinates where the feature is recorded, but in a larger area. For example, a 311-pothole recorded in a three-way intersection will probably affect cycling safety perception in all three street segments. To account for that, we create a radius buffer of $rb = 5m$ such that any social feature event recorded will be counted towards all street segments covered by a radius of five meters around its own geolocation. For the crime rates feature, we enlarge that radius buffer to $rb = 500m$ since crimes can potentially have larger areas of influence [26].

B. Ground Truth Collection

We launched the rating platform with cycling videos for the city of Washington D.C. The platform was promoted by Washington's Area Cyclist Association (WABA) as well as by several other smaller cyclist associations through cycling events, blog posts and social media feeds to encourage cyclists to access the website and rate the safety of as many cycling videos as they could. For this paper, we use all the video safety ratings collected over a period of three months. We collected $1,476$ ratings from 159 different participants, covering a total of 443 city street segments. Each segment safety label $L_i \in$ [1-5] was computed after averaging all the collected individual ratings across video coverage and participants, and assigning it to its closest integer value. We had an average of 5.1 ratings per street segment, and these provided perceived safety information for over 3% of all the street segments in the city of D.C. (out of a total of $13,462$). The ratings collected covered, proportionally, the 13 different road types in Washington's D.C. road network. Of the total 159 participants, 42.5% of the participants self-declared themselves as fearless and 45.4% as confident, followed by 9.2% interested in cycling and 2.9% reluctant.

C. Methods

We create the training and testing dataset as a set with all the 443 street segments and their perceived segment safety labels computed for the city of Washington D.C. This unique dataset will be shared as an open resource for researchers and practitioners working in transportation-related analyses. Each street segment is characterized by either $|F_i| = 69$ or $|F_i| = 211$ different built-in environment and social features depending on whether the social features are measured by total volumes or by volumes per type, as explained in the previous section. We evaluate the classification accuracy of the segment safety levels using the following battery of methods: Support Vector Machines (SVM), Decision Trees (DT), Bagging for DTs (BAG), Random Forest (RF), Gradient Boosting (GBoost) and Extreme Gradient Boosting (XGBoost); and compare all these techniques against a simple baseline that considers all safety labels in our dataset to be the majority label. Finally, we also evaluate the impact that sparsity and class imbalance have on the accuracy of the methods.

We evaluate each method with the following sets of features: (a) *built-in environment features only*, which applies the prediction methods over a training dataset that contains the perceived safety labels and only built-in features as predictors (*BuiltEnv*); (b) *social features (total) only*, which applies the prediction methods over street segments characterized by their perceived safety labels and social features only, computed using the monthly average across types approach (*Social[total]*); (c) *social network features (type) only*, as above, but computing the social network features as monthly average per type, thus increasing the size of the predictive feature vector considered from 69 to 211 (*Social[type]*); (d) *built-in environment and social features (total)*, which applies the prediction methods over street segments characterized by both built-in environment

Table I
MICRO-F1 (M-F1) AND MACRO-F1 (M-F1) SCORES FOR EACH METHOD (ROWS) AND SET OF FEATURES (COLUMNS).

METHOD / FEATURES	BuiltEnv	Social [total]	Social [type]	BuiltEnv+Social [total]	BuiltEnv+Social [type]
SVM	0.59/0.31	0.52/0.27	0.54/0.31	0.58/0.34	0.58/0.36
Decision Trees (DT)	0.46/0.34	0.48/0.26	0.49/0.30	0.56/0.31	0.52/0.36
Bagging DT (BAG)	0.60/0.43	0.52/0.29	0.57/0.40	0.62/0.36	**0.65/0.42**
Random Forest (RF)	**0.62/0.45**	0.54/0.30	0.57/0.39	0.63/0.37	0.63/0.41
Gradient Boosting (GBoost)	0.60/0.41	0.55/0.31	0.58/0.41	0.62/0.40	0.64/0.44
XGBoost	0.57/0.37	0.55/0.34	**0.59/0.43**	0.62/0.37	**0.65/0.44**
Baseline	0.45/0.13	0.45/0.13	0.45/0.13	0.45/0.13	0.45/0.13

and social features represented using the monthly average approach across all types; and (e) *built-in environment and social features (type)*, as above, but with social features computed using the monthly average per type approach. These analyses, together with different classification methods, will aid in the evaluation of how the sparsity of the feature vectors affects the classification accuracy.

We divide the dataset into 80-20% random splits of training and testing subsets, repeat this process 10 times and report average safety level prediction accuracies for each method a set of features. To account for the effect of the imbalanced nature of our dataset, we report and analyze both micro- and macro-F1 scores. Significantly lower micro scores when compared to macro values, reflect high misclassification among the most common labels, with labels with lower numbers of samples being correctly classified. On the other hand, macro scores significantly lower than micro scores are associated to poor classification rates among labels with lower numbers of samples, with common labels being correctly classified. Table I shows the main results for each method and set of features. The overall trend shows that, in general, considering only built-in environment features yields slightly better results than considering only social features (maximum micro F-1 of $m\text{-}F1 = 0.62$ vs. $m\text{-}F1 = 0.59$); and that both results are between 14-17% better that the majority vote baseline (with a $m\text{-}F1 = 0.45$). This result reveals that variables such as the type of road, slope, the centrality of the street segment, or the presence of biking facilities are by themselves more predictive of perceived cycling safety than variables that characterize the social environment such as crashes, crime rates, 311 bicycle-related complaints or parking and moving violations. We hypothesize that this could be due to the fact that built-in environment features are directly experienced by cyclists every time they travel, while the social features require awareness about the events that happen in the streets. In other words, for social features to have an impact on the prediction of cycling safety levels, cyclists need to be acquainted with their environment, which might not always be the case unless they are informed citizens or familiar with the area they are cycling. For example, a cyclist going through a street might not know that crime rates in that area are high, or that the street has one of the highest indices in bicycle crashes in the city. However, as she is cycling through the street, she will directly perceive the slope or the presence of a bike lane. Interestingly, previous work based mostly on surveys and qualitative studies has also shown that while certain variables of the physical environment such as slope or centrality are highly statistically significantly related to cycling safety ($p < 0.001$), social features such as crime, have more marginally significant associations ($p < 0.10$) [27], [28], [29].

These results highlight similar findings to ours although using different analytical approaches and focusing on relationships between features rather than on prediction. Importantly, these results also show that by exclusively using a few features extracted from Open Street Map and Google's Elevation Service, the safety level accuracy rates can be quite high ($m\text{-}F1 = 0.62$), indicating that for those cities that do not have the resources to collect any other type of open data, there is still an opportunity to extract fairly good cycling safety maps exclusively using information from open and crowdsourced platforms. Although built-in environment features are highly predictive of cycling safety by themselves, the most accurate predictions are obtained when combining built-in environment and social features. Furthermore, considering average feature values per type yields better results than averages across types, which reflects that albeit being sparser, the by type feature vectors are also more informative thus improving prediction rates. For cities with the ability to collect social features, this reveals that the prediction of cycling safety ratings can be improved with F1-scores between 3-8% higher depending on the method and set of features used.

Table I shows that the best result was obtained with XGBoost using segments characterized with both built-in environment features and social features represented by type, which lead to $m\text{-}F1 = 0.65, M\text{-}F1 = 0.44$; followed by Bagging using segments characterized with both built-in environment features and social features represented by type, which lead to $m\text{-}F1 = 0.65, M\text{-}F1 = 0.42$. Interestingly, XGBoost has been reported to work well with sparse feature vectors [30]. These prediction approaches improved the majority vote baseline by 20%. The most predictive features identified by XGBoost and ordered by feature importance im and their standard deviation std included, centrality of the street segment ($im = 0.02, std = 0.001$), presence of cycling facilities ($im = 0.019, std = 0.001$), crime

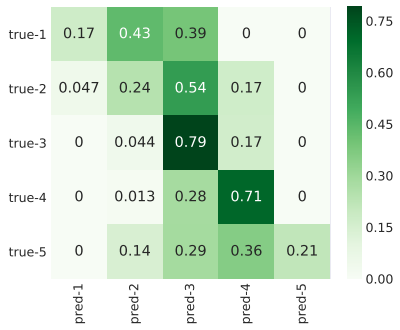

Figure 1. Confusion matrix for XGBoost. Percentage of true (rows) versus predicted (columns) values.

rates (theft or burglary) ($im = 0.018, std = 0.001$) and slope ($im = 0.017, std = 0.001$). Looking at the average values of these features for each cycling safety level showed that cyclists associate higher safety ratings to highly central segments that have many cycling facilities, low crime rates and low slope. We also observe that XGBoost has micro-F1 scores slightly higher than the macro-F1 scores which indicates that the classification rates among labels with lower number of samples (cycling safety labels 1, 2 and 5) are associated to poorer results than the most common labels.

Fig. 1 shows the confusion matrix for the best approach, XGBoost, averaged over 10 runs with different 80-20% dataset partitions. The values in each matrix row $true$-i with $i \in [1-5]$ represent the average percentage of samples with true label i classified as $pred$-j with $j \in [1-5]$ across all runs. We can observe that (1) the most common labels (3 and 4) are for the most part correctly predicted as its own label (79% and 71%, respectively); and that (2) for the labels with lower number of samples (1, 2 and 5), the largest percentage of wrongly predicted samples is always one safety level away from the correct one *e.g.,* for safety level 1 ($true$-1), 43% of the samples are predicted as safety level 2; while for safety level 2 ($true$-2), 54% of the samples are predicted as safety level 3. A similar matrix was obtained for the second best approach, Bagging over Decision Trees. This finding is highly relevant from a cycling safety policy perspective since decision makers will be able to assert that although the accuracy of the predictions is not perfect, in a large number of cases the incorrect prediction will be next to its correct label rather than few levels away. If cycling safety maps are used to identify areas that require immediate action, the fact that the incorrect labels are just one step away from its true values guarantees that the decisions will be adequate. Similarly, for cyclists using these maps to choose cycling routes, the selected streets would be not far off from their expected cycling safety level. In the next two sections, we explore two different approaches to improve the F1 scores: (a) address the imbalanced nature of the dataset and (b) take into account the experience of the participants that provide ratings.

1) Imbalanced Dataset.: Given that the nature of our dataset is imbalanced *i.e.,* we have more samples with safety values 3 and 4 than any other labels, we also evaluate two different approaches to potentially improve the F1 scores. First, we explore the use of over and undersampling techniques in combination with feature selection techniques; and second, we evaluate the use of only three or four segment cycling safety levels instead of five *i.e.,* video ratings are transformed into street segment labels L_i as explained in section III-B, but scaled to ranges [1-3] or [1-4] instead of [1-5]. Although this approach decreases the granularity of the safety ratings provided, it could be justified if the F1 scores are much higher, since it would provide more accurate cycling safety maps. We first focus on over and undersampling. Undersampling reduces the number of samples of each class to the smallest value, and repeats the process several times to account for selection biases. On the other hand, oversampling creates synthetic samples, via k-nearest neighbors, for all classes until they reach the number of samples in the majority class. We used SMOTE to implement both methods and the resulting F1 scores are shown in Table II. For both over and undersampling we also evaluated the use of a feature selection technique prior to the execution of SMOTE. Specifically, we considered mRMR and recursive feature elimination with cross-validation (RFECV). Additionally, for over-sampling, we evaluated both regular-SMOTE and SVM-SMOTE. As Table II shows, oversampling slightly improved the XGBoost classifier by 1% when no feature selection and a regular SMOTE were used over both built-in environment and social features (by type).

On the other hand, we also re-run all methods and sets of features considering only three or four segment safety levels instead of five. Table II (bottom) shows the results for the best methods. As expected, reducing the number of cycling safety levels improved the F1 scores. Considering only three cycling safety levels improved the best F1 score by 21% with micro and macro scores of m-$F1 = 0.87, M$-$F1 = 0.54$. Importantly, this result was also better than the majority vote baseline when only three classes are considered (m-$F1 = 0.78, M$-$F1 = 0.3$). Similarly, when considering four safety levels instead of five, the best approach slightly improved the F1 scores obtained with five safety levels by 2% (m-$F1 = 0.67, M$-$F1 = 0.49$); and it also improved its 4-class majority baseline (m-$F1 = 0.54, M$-$F1 = 0.17$). These 3- and 4-safety level experiments offer cities and cyclists the possibility of reaching higher accuracy in the prediction of cycling safety by sacrificing the number of levels considered. The total number of cycling safety levels used highly depends on the corresponding departments of transportation and decision makers; for example while Semler *et al.* used four levels, Sorton *et al.* presented analyses with five [31], [11]. It is up to the users of the system, city planners or cyclists, to evaluate whether three, four or five cycling safety levels are the ideal approach for their decision

Table II
MICRO-F1 (M-F1) AND MACRO-F1 (M-F1) SCORES WHEN CLASS IMBALANCE IS ADDRESSED WITH OVER/UNDERSAMPLING.

METHOD / FEATURES	BuiltEnv	Social [total]	Social [type]	BuiltEnv+Social [total]	BuiltEnv+Social [type]
Oversampling XGBoost (none, regular)	0.60/0.37	0.52/0.33	0.60/0.40	0.63/0.41	**0.66/0.44**
Oversampling BAG (mRMR,SVM)	0.60/0.42	0.51/0.29	0.61/0.40	0.62/0.36	**0.65/0.42**
Undersampling XGBoost (mRMR)	0.31/0.21	0.23/0.20	0.26/0.22	0.31/0.21	0.32/0.24
Undersampling BAG (mRMR)	0.32/0.21	0.24/0.20	0.27/0.23	0.31/0.23	0.30/0.24
Three-Class (XGBoost)	0.84/0.43	0.84/0.33	0.87/0.53	0.85/0.43	**0.87/0.54**
Four-Class (GBoost)	0.66/0.41	0.58/0.30	0.63/0.35	0.66/0.42	**0.67/0.49**

making and cycling route choice processes.

2) Experience and Familiarity.: The evaluation presented so far assumes that all safety ratings provided by participants are equally informative *i.e.,* each street segment i is assigned a cycling safety label L_i computed as the closest integer in [1-5] to the average of all ratings collected for that segment (see *Default* in Table III, this formula has been simplified for clarity purposes; recall that the ratings are also weighted by the percentage of the segment covered across all videos as explained in section III-B). However, a better approach might be to give more importance to the safety ratings provided by participants that recognize and are familiar with the cycling route shown in the video, or by participants who have a certain level of cycling experience. The crowdsourced rating platform that we have developed gathers information about the participants both in terms of cycling experience as well as of the familiarity with the cycling route whose safety level is being labeled. Recall that participants are asked about their cycling experience via a screen where four options are offered: fearless, confident, interested or reluctant. On the other hand, every time a participant rates a video, she is also asked about her familiarity with the route by choosing between two options: familiar or not familiar (although participants can also leave this selection blank).

In this section, we explore the use of the cycling experience and familiarity features to change the relevance of the individual safety ratings (and, in turn, the safety labels assigned to each street segment) and evaluate its impact on the accuracy of the safety level prediction. Specifically, we evaluate the following weighting schemes (see Formula column in Table III). The *familiarity* scheme assumes that cyclists familiar with the route they are rating have probably cycled it multiple times thus gaining a more accurate understanding of its safety level besides the observable information provided in the recorded cycling video. Thus, the ratings provided by cyclists who are familiar with the route are weighted higher than those from cyclists who are not familiar with the route or who have not provided a familiarity score. Table III shows the formula for this scheme, where individual ratings r_j are weighted by familiarity with $FAM = 3$ being familiar with the route, $FAM = 1$ being not familiar and $FAM = 2$ is the weight assigned when no familiarity information has been provided *i.e.,* we do not know whether the participant is acquainted or not with the route. The next three schemes, explore various approaches using the cycling experience of

Table III
M-F1/M-F1 SCORES PER SCHEME FOR 4 SAFETY LEVELS, FAM: FAMILIARITY AND E: CYCLING EXPERIENCE.

Scheme	Formula	F1 Scores
Default	$\sum_{j=1}^{n} r_j/n$	0.67/0.49
Familiarity	$\sum_{j=1}^{n} FAM_j r_j / \sum_{j=1}^{n} FAM_j$	**0.69/0.51**
Fearless	$\sum_{j=1}^{n} E_j r_j / \sum_{j=1}^{n} E_j$	0.68/0.50
Reluctant	$\sum_{j=1}^{n} (5-E_j) r_j) / \sum_{j=1}^{n} (5-E_j)$	0.67/0.46
Experience	$(E_j = 4 \& r_j < 3) \rightarrow r_j -= 1$	**0.68/0.55**
	$(E_j = 1 \& r_j > 3) \rightarrow r_j += 1$	
	$(E_j = 3 \& r_j < 3) \rightarrow r_j -= .5$	
	$(E_j = 2 \& r_j > 3) \rightarrow r_j += .5$	

the participants that rate the videos. Scheme two, *fearless*, argues that fearless cyclists tend to provide the most accurate ratings, while reluctant cyclists are assumed to be the most conservative in their labeling. Thus, the safety label for each segment is re-computed giving more importance to ratings provided by cyclists with higher experience. Table III shows the formula for the fearless scheme with experience $E = 1$ being a reluctant cyclist, 2 interested, 3 confident and 4 fearless. The third scheme, *reluctant*, assumes exactly the opposite of the previous scheme, that is, that reluctant cyclists are the ones that provide the most accurate ratings and that, on the other extreme, fearless cyclists tend to be too optimistic and label everything as highly safe. To reflect this, the safety labels for each street segment are modified as shown in Table III, giving the highest weight to ratings provided by reluctant cyclists ($E = 1$), followed by interested, confident and fearless ($E = 4$). Finally, the fourth scheme, *experience*, is a combination of the last two schemes that works under the hypothesis that while reluctant cyclists (and confident cyclists, to a lesser degree) might be the best at identifying highly safe street segments (safety ratings $r > 3$), fearless cyclists (and interested cyclists, to a lesser degree) might the best at pinpointing dangerous streets (safety ratings $r < 3$). In this case, since ratings will be modified differently based on cyclist experience, we define rules for each experience level as shown in Table III. For fearless cyclists, dangerous safety ratings are re-scored as even more dangerous (decrease rating by one) while for reluctant cyclists, safe ratings are re-scored as safer (increase rating by one). Similar re-scoring is applied to confident and interested cyclists respectively, although the rating modification is smaller (.5).

We apply each of these weighting schemes to the collected cycling safety ratings, re-computing the segment safety labels L_i, and replicate the safety level prediction experiments from the previous section considering three, four or five segment safety levels *i.e.,* $L_i \in$ [1-3], [1-4] or [1-5]. The third column in Table III shows the micro- and macro-F1 scores when four cycling safety levels are considered. We can observe that all weighting approaches, except for reluctant, improved both the micro and macro scores of the default case which does not take into account either familiarity or cycling experience. There were two top results: the familiarity scheme improved by $\approx 2\%$ both the micro- and macro-F1 scores; while the the experience scheme improved the micro score by 1% and boosted the macro-F1 score by 6%. Considering five or three different cycling safety levels boosted the default macro-F1 scores by 3% for all scenarios, without improving the micro F1 scores. The best improvements with five classes were obtained using the experience scheme and with three classes using the familiarity scheme.

V. Conclusions

We have presented an approach to predict urban cycling safety at the street segment level exclusively using information from open and crowdsourced datasets. Our evaluation for the city of Washington D.C. shows that a combination of built-in environment and social features, extracted from D.C.'s open data portal and from Open Street Maps, provides F1-scores of up to 88%. Additionally, we have also shown that taking into account class imbalance or cycling experience slightly increases the accuracy of the predictions.

References

[1] G. Lindsay, A. Macmillan, and A. Woodward, "Moving urban trips from cars to bicycles: impact on health and emissions," *Australian and New Zealand Journal of Public Health*, vol. 35, no. 1, pp. 54–60, 2011.

[2] D. R. Bassett, J. Pucher Jr, R. Buehler, D. L. Thompson, and S. E. Crouter, "Walking, cycling, and obesity rates in europe, north america, and australia," *Journal of Physical Activity and Health*, vol. 5, no. 6, pp. 795–814, 2008.

[3] R. Cervero, "Progressive transport and the poor: Bogota's bold steps forward," *Access Magazine*, vol. 1, no. 27, 2005.

[4] CDC, "Center for disease control and prevention." https://www.cdc.gov/injury/wisqars/index.html, 2015.

[5] B. Landis, V. Vattikuti, and M. Brannick, "Real-time human perceptions: toward a bicycle level of service," *TRB Journal*, no. 1578, pp. 119–126, 1997.

[6] M. A. Figliozzi and B. P. Blanc, "Evaluating the use of crowdsourcing as a data collection method for bicycle performance measures," *Portland State University, TR 768*, 2015.

[7] M. Møller and T. Hels, "Cyclists' perception of risk in roundabouts," *Accident Analysis & Prevention*, vol. 40, no. 3, pp. 1055–1062, 2008.

[8] OpenDataSoft, https://www.opendatasoft.com/a-comprehensive-list-of-all-open-data-portals-around-the-world/, 2018, [accessed July-30-2018].

[9] M. Haklay and P. Weber, "Openstreetmap: User-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, 2008.

[10] W. Davis, "Bicycle safety evaluation," Ph.D. dissertation, Auburn University, 1987.

[11] A. Sorton and T. Walsh, "Bicycle stress level to evaluate urban and suburban bicycle compatibility," *TRR*, 1994.

[12] J. Bolitho, "A multi-stage, multi-faceted approach to addressing'car dooring'in inner melbourne," in *Australasian College of Road Safety Conference*, 2013.

[13] J. Kerr and et al., "Perceived neighborhood environmental attributes associated with walking and cycling," *Environmental Health Perspectives*, vol. 124, no. 3, p. 290, 2016.

[14] L. W. Sherman, "Hot spots of crime and criminal careers of places," *Crime and Place*, vol. 4, pp. 35–52, 1995.

[15] P. Schepers and M. e. a. Hagenzieker, "A conceptual framework for road safety," *Accident Analysis & Prevention*, vol. 62, pp. 331–340, 2014.

[16] M. Dozza and J. Werneke, "Introducing naturalistic cycling data," *Transportation Research Part F*, vol. 24, 2014.

[17] J. Parkin and et al., "Models of perceived cycling risk and route acceptability," *Accident Analysis & Prevention*, vol. 39, no. 2, pp. 364–371, 2007.

[18] M. Klobucar and J. Fricker, "Network evaluation tool to improve real and perceived bicycle safety," *TRB Journal*, no. 2031, pp. 25–33, 2007.

[19] C. Pooley and M. e. a. Tight, *Understanding walking and cycling: Summary of key findings and recommendations*.

[20] J. Leskovec and R. Sosič, "Snap: A general-purpose network analysis and graph-mining library," *ACM TIST*, vol. 8, 2016.

[21] S. Porta, P. Crucitti, and V. Latora, "The network analysis of urban streets: a dual approach," *Physica A*, vol. 369, 2006.

[22] ——, "The network analysis of urban streets: a primal approach," *Environment and Planning B*, vol. 33, 2006.

[23] J. Dill and N. McNeil, "Four types of cyclists?" *Transportation Research Record: Journal of the Transportation Research Board*, no. 2387, pp. 129–138, 2013.

[24] Mapbox, "Map matching api," www.mapbox.com, 2018, [accessed July-30-2018].

[25] P. Newson and J. Krumm, "Hidden markov map matching through noise and sparseness," in *ACM SIGSPATIAL*, 2009.

[26] I. Janssen, "Crime and perceptions of safety in the home neighborhood," *Preventive Medicine*, vol. 66, 2014.

[27] J. Broach, J. Gliebe, and J. Dill, "Bicycle route choice model developed using revealed preference gps data," in *TRB Meeting*, 2011.

[28] P. Crucitti, V. Latora, and S. Porta, "Centrality in networks of urban streets," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 16, no. 1, 2006.

[29] F. De Meester and et al., "Does the perception of neighborhood built environmental attributes influence active transport in adolescents?" *International Journal of Behavioral Nutrition and Physical Activity*, vol. 10, no. 1, 2013.

[30] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *SIGKDD*. ACM, 2016, pp. 785–794.

[31] C. Semler and et al., "Low-stress lts: District of columbia's innovative approach to applying level of traffic stress," in *TRB Meeting*, 2017.