Open and Crowdsourced Data to Predict and Characterize Perceived Cycling Safety

Vanessa Frias-Martinez
iSchool and UMIACS
University of Maryland
The Benefits of Cycling
Cycling Safety is a Concern

• ~500K cyclists injured (~700 deaths) in 2013 (CDC Injury Center)

• Increase in both number of people riding and number of deaths
Causes behind Cycling Accidents

• Inadequate infrastructure
• More drivers on the road (low gas prices)
• Smartphone use and distractions
• Increasing population in urban areas
Approaches to Increase Safety

• Vision Zero initiatives to eliminate all traffic fatalities include:
  – Proactive policy
  – Infrastructure changes
  – Education

• Initiatives have not always been successful
  – In 2018 - LA 5% increase in cyclist and pedestrian deaths
Understanding Safety Perception at the Street Level

• Safety measures focus a lot on crash numbers, which is an incomplete statistic
• We need a better understanding of perceived cycling safety at the street level
Understanding Safety Perception at the Street Level

- Identify locations where changes might be more needed (decision makers, cyclists and advocacy groups)
Understanding Safety Perception at the Street Level

• Identify locations where changes might be more needed (decision makers, cyclists and advocacy groups)

• Evaluate connectivity and cycling safety per community to reveal accessibility and equity issues
Cycling Safety Maps
Cycling Safety Maps

- Associations between Attributes and Cycling Safety Perceptions
Attributes

• Measures: traffic speed, traffic volume, frequency of parking turnover
  – Require expensive sensors that cannot be available in every street

• Observations from video recordings
  – Expensive and not scalable
Cycling Safety Perceptions

• Cycling safety perceptions associated to attributes are based on:
  – Logical intuitions (e.g., more cars, less safe)
  – Qualitative studies, generalizability not validated
Proposed Approach - Attributes

Can we find more affordable and scalable attributes for cycling safety?
Proposed Approach – Perception Associations

Can we formally validate that the attributes are predictive of cycling safety perception?
New Approach to Perceived Cycling Safety Maps
Explore the use of Open Datasets and Open Street Maps as a source for perceived cycling safety attributes
Open Data

• Lowering the bar to comprehensive cycling safety maps:
  – Open Data Repositories: 2600 cities worldwide (some cities have the data, but not public)
  – Open Street Maps: 4 million small- to mid-sized cities
Our Approach

Crowdsourcing cycling safety perceptions from cyclists (ground truth) and build a ML model to test associations between attributes and safety perceptions.
Predicting Perceived Cycling Safety Levels Using Open and Crowdsourced Data

Fig. 1. Cycling Safety Prediction. The proposed approach consists of three main components: (a) the extraction of cycling safety levels prediction features from open data portals and Open Street Map, (b) an open source, crowdsourced rating platform to collect the ground truth cycling safety labels necessary to accurately train the cycling safety prediction models, and (c) the development and evaluation of accurate and transparent cycling safety prediction models.

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 [36]. 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, passing stop sign without coming to full stop 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 of a given street segment.

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 [24, 27]. 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 Map visualization figures in this paper are created using Lea et al. and Carto tiles.
A. Perceived Cycling Safety Attributes

Framework

1. Prediction Features
   - Open Data Portal
   - Open Street Map

2. Segment-level Feature Extraction

3. Feature Maps

4. Classification Models

5. Evaluation & Application

A.
Perceived Cycling Safety Attributes

- Record Cycling Videos with GPS
- Video-Segment Mapping
- Mapbox's Map Matching API
- Rating Collection
- Ground Truth Map
- Predicted Map

- 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;
- 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;
- Parking volumes have been shown to impact safety perception, with higher parking volumes associated to less safety. 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, passing stop sign without coming to full stop 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 of a given street segment.

- 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:
  - 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.
Predicting Perceived Cycling Safety Levels Using Open and Crowdsourced Data

Fig. 1. Cycling Safety Prediction. The proposed approach consists of three main components: (a) the extraction of cycling safety levels prediction features from open data portals and Open Street Map, (b) an open source, crowdsourced rating platform to collect the ground truth cycling safety labels necessary to accurately train the cycling safety prediction models, and (c) the development and evaluation of accurate and transparent cycling safety prediction models.

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; and (e) parking volumes have been shown to impact safety perception, with higher parking volumes associated to less safety. 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, passing stop sign without coming to full stop, 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 of a given street segment.

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; (g) graph-based Map visualization figures in this paper are created using Lea et al Carto tiles.

B. Ground Truth (Validation Data)
Predicting Perceived Cycling Safety Levels Using Open and Crowdsourced Data

Fig. 1. Cycling Safety Prediction. The proposed approach consists of three main components: (a) the extraction of cycling safety levels prediction features from open data portals and Open Street Map, (b) an open source, crowdsourced rating platform to collect the ground truth cycling safety labels necessary to accurately train the cycling safety prediction models, and (c) the development and evaluation of accurate and transparent cycling safety prediction models.

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 [36]. 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, passing stop sign without coming to full stop 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 of a given street segment.

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 [24, 27]. 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 Map visualization figures in this paper are created using Lea et al. and Carto tiles; (h) {

Predicting Perceived Cycling Safety Levels Using Open and Crowdsourced Data

The proposed approach consists of three main components: (a) the extraction of cycling safety levels prediction features from open data portals and Open Street Map, (b) an open source, crowdsourced rating platform to collect the ground truth cycling safety labels necessary to accurately train the cycling safety prediction models, and (c) the development and evaluation of accurate and transparent cycling safety prediction models.

Crash statistics per street segment are typically available by total volume or by type of crash, e.g., collision with a 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; and finally, (e) parking volumes have been shown to impact safety perception, with higher parking volumes associated to less safety. 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, passing stop sign without coming to full stop, 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 of a given street segment.

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. 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 Map visualization - figures in this paper are created using Lea and Carto tiles; (h)...

Framework

A. Perceived Cycling Safety Attributes

B. Ground Truth (Validation Data)

C. Perceived Safety Prediction

Evaluation & Application

Classification Models

Predicted Map

Feature Maps

Ground Truth Map

Rating Collection

Video-Segment Mapping

Mapbox's Map Matching API

Record Cycling Videos with GPS

Crowdsourced Rating Platform

Segment-level Feature Extraction

Open Street Map

Open Data Portal

Prediction Features
A. Perceived Cycling Safety Attributes
Qualitative research on cycling safety factors has identified that these factors play a role in safety perception:

- Social fabric e.g., crime rates (Open Datasets)
- Built environment e.g., presence of cycling facilities (Open Street Maps)
Social Attributes

• Crime rates
• Points of interest
• Bicycle crashes
• 311 requests related to street conditions
• Parking and moving violations
Impact Buffer

Built-in Environment Feature: Cycling Facilities retrieved from Open Street Map

Social Feature: 311 Pothole Requests. Example of the buffering approach.

Fig. 4. Examples of built-in environment (a) and social features (b) for the area of Columbia Heights in Washington, D.C.

Cycling facilities are extracted from Open Street Map and assigned to street segments based on coverage. Social features are assigned to street segments within a given radius. In (b), all 311 pothole requests for 2017 are shown in the large map. Zooming in, we observe how requests a, b, c and d are counted towards all street segments covered by a 5m radius: only one street segment for a and c, four street segments for b, or none for d.

Using this approach, 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 social features is an issue for the perceived cycling safety predictions.
Built Environment Attributes

- Road network characteristics
- Presence of cycling facilities
- Graph-based road network features
Graph-based Features

Fig. 2. Transformation of the road network defined by two blocks in Washington, D.C., into its undirected primal and dual graphs. The road network consists of street segments a to g. Given the graphs, graph-based features that characterize the centrality of the street segments are computed using the SNAP package.

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 [41, 42]. 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) [28]. 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 [43, 44] (see Figure 2 for an example); 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 as discussed in Section 2 [32, 36, 37, 50].

3.2 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 level [1]. 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. Although one might argue that cities could use predictive models trained for other cities, thus eliminating the need to collect ground truth data, it is highly probable that models trained in other cities will not be able to capture the local conditions well, thus decreasing the prediction accuracy rates. In fact, related literature has shown that the impact of the social features and, to a lesser extent, of the built-in environment features in cycling safety perception changes across cities, countries and cultures [8, 51]. Finally, the collection of ground truth safety labels only needs to be executed once, after which safety predictions will change as the prediction features change over time.

Figure 3 shows consecutive snapshots of the platform, as the user navigates through the different steps. After logging in (step 1), users are asked to rate their cycling experience level by choosing among the four following options: fearless, confident, interested or reluctant (see step 2 in Figure 3). These four types of cyclists are based on the taxonomy created by Geller et al. and largely used in the cycling literature [14, 15]. Next, we collect survey information with respect to cycling, demographic and socio-economic characteristics of the user including type of biking (e.g., utility or recreational), gender, educational level or income, among others (see Personal Features...
Attributes for DC

• 63 built environment features
  – 11 road network types
  – 39 graph-based (centrality measures)
  – 13 cycling facilities types

• Social features: monthly average across types (6) and monthly average per type (148)
  – 11 types for crime data
  – 11 types for crash data
  – 72 for 311 requests
  – 10 POIs
  – 36 types of parking violations
  – 8 moving violations
B. Ground Truth Data Collection
Predicting Perceived Cycling Safety Levels Using Open and Crowdsourced Data

A. Perceived Cycling Safety Attributes

1. 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.

2. 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.

3. Parking volumes have been shown to impact safety perception, with higher parking volumes associated to less safety. 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, passing stop sign without coming to full stop, 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 of a given street segment.

B. Ground Truth (Validation Data)

C. Perceived Safety Prediction

1. On the other hand, we explore the following built-in environment features as potential predictive proxies for cycling safety at the segment level: 
   a. 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. 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.

2. Graph-based Map visualization figures in this paper are created using Lea and Carto tiles.
Ground Truth Collection

- Recorded cycling videos in Washington, D.C
- Built a webpage to crowdsourced cycling safety perceptions
- WABA promoted our project in cycling events
- Collected cycling safety perceptions from cyclists
Crowdsourced Safety Perceptions

1. Login Page
2. Choose Experience Level
3. Survey (optional)
4. Cycling Safety Rating
5. Cycling Safety Map

Fig. 3. Crowdsourced rating platform: description of the steps participants go through. First, users are asked to login and provide their cycling experience level, followed by an optional survey. A user that does not feel comfortable sharing personal data. After the survey, the user will be shown 20 cycling videos recorded by actual cyclists and after each video, she will be prompted to provide a cycling safety rating between 1 (too dangerous, I would never ride there) and 5 (very safe, even a kid could ride there); as well as to choose among different reasons as to why such rating was provided, including traffic, bike lane design or driving quality, among others (see step 4 in the Figure). Users can select one or multiple reasons per safety rating provided. Table 1 shows the complete list of safety ratings and rating reasons that the cyclists can select from. Additionally, 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. Platform users can watch and rate as many videos as they want, and videos are shown randomly for the first time. However, once a video has at least one rating the probability of being shown again to another platform user will be slightly higher so as to guarantee that a cycling safety level for a segment is not exclusively based on one individual rating.

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 straightforward since GPS sensors have errors, and more so in urban environments where when surrounded by tall buildings the GPS might lose 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. 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.
Cycling Safety Tool
From Videos to Segments

- Videos are rated multiple times by cyclists
- Each segment might appear in multiple videos
- Final segment label (1-5) is averaged across video ratings and weighted by % of street segment present in video
## Personal and Rating Features

<table>
<thead>
<tr>
<th>Personal Features</th>
<th>Safety Ratings</th>
<th>Rating Reasons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Usual trip purpose</td>
<td>1: too dangerous, I would never ride there</td>
<td>Traffic</td>
</tr>
<tr>
<td>Age</td>
<td>2: a bit dangerous, I wouldn’t ride here unless I have to</td>
<td>Bike lane design (or lack of)</td>
</tr>
<tr>
<td>Ethnicity</td>
<td>3: fair, I need to be cautious to ride here</td>
<td>Bike lane blocked (vehicle)</td>
</tr>
<tr>
<td>Education level</td>
<td>4: quite safe, I would easily ride here</td>
<td>Dooring (car door might hit cyclist)</td>
</tr>
<tr>
<td>Marital status</td>
<td>5: very safe, even a kid could ride here</td>
<td>Pedestrians crossing</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td>Intersection design</td>
</tr>
<tr>
<td>Driver’s license</td>
<td></td>
<td>Driving quality</td>
</tr>
<tr>
<td>Access to car</td>
<td></td>
<td>Road quality (paving)</td>
</tr>
<tr>
<td>Household income</td>
<td></td>
<td>Hill</td>
</tr>
<tr>
<td>Length of residence in city</td>
<td></td>
<td>Neighborhood security</td>
</tr>
<tr>
<td>Type of biking</td>
<td></td>
<td>Weather</td>
</tr>
</tbody>
</table>
Ground Truth Collection

1476 ratings from 159 participants
Ground Truth Collection

Average rating = very dangerous (1).

Average rating = very safe (5).

Fig. 6. Snapshot of two videos shown in the rating platform. The video on the left had an average rating of one: very dangerous, while the video on the right was rated on average as very safe (rating of five). The reasons provided to justify such ratings were traffic, bike lane design and dooring for the very dangerous video, and bike lane design for the very safe video.

14% in the 45-54 range, followed by all other age groups with percentages smaller than 7%.

As for gender, the distribution was almost even with 48% males and 43% females (the remaining 9% did not provide any gender information).

Finally, 41.5% of the participants self-declared themselves as fearless and 44% as confident, followed by 8.2% interested in cycling and 1.9% reluctant (the remaining 4.4% did not answer this question).

On the other hand, Figure 5 shows some statistics about the video safety ratings provided by participants and the reasons behind them. Figure 5(a) shows the distribution of number of ratings per participant. We observe that a high percentage of participants (64%) watched and rated 1-10 cycling videos; 23% of participants provided 10-20 ratings; 6% watched 20-30 videos and the remaining 7% were highly active participants rating between 30 and 70 videos each. In Figure 5(b), we observe that the distribution of video ratings per cycling safety level follows a normal distribution as observed in many other rating tasks with multiple values \[25\]; with a large percentage (over 60%) of average cycling safety levels (ratings 3 and 4) and with smaller percentages (5-12%) of more extreme levels (ratings 1, 2 and 5). Such distribution was also observed for the segment cycling safety levels. Finally, Figure 5(c) shows that the most common factors affecting the rating were the bike lane design (66%), traffic (51%), dooring (32%) and road quality (25%). Recall these were multiple-choice and participants could select more than one feature per rating. As an example, Figure 6 shows the snapshots of two videos shown in the rating platform. The video in (a) had an average rating of very dangerous, with the most common reasons being traffic, bike lane design and dooring (car door might hit cyclist); while the video in (b) was rated, on average, as very safe, mostly due to bike lane design.

4.3 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) \[39\] and Extreme Gradient Boosting (XGBoost) \[10\]; 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, spatial autocorrelation and class imbalance have on the accuracy of the methods.
C. Perceived Cycling Safety Prediction
Predicting Perceived Cycling Safety Levels Using Open and Crowdsourced Data

Fig. 1. Cycling Safety Prediction. The proposed approach consists of three main components: (a) the extraction of cycling safety levels prediction features from open data portals and Open Street Map, (b) an open source, crowdsourced rating platform to collect the ground truth cycling safety labels necessary to accurately train the cycling safety prediction models, and (c) the development and evaluation of accurate and transparent cycling safety prediction models.

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 [36]. 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, passing stop sign without coming to full stop 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 of a given street segment.

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 [24, 27]. 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 Map visualization figures in this paper are created using Lea et al. and Carto tiles. 

Perceived Cycling Safety Prediction

• To assess whether open and crowdsourced data can be used to
  – predict perceived cycling safety
  – assess associations between attributes and cycling safety perceptions
Prediction Results

• Dataset:
  – Segments with features
  – Crowdsourced cycling safety labels
  – mRMR feature selection
  – 70%-30% training-testing 10 times and report averages
## Prediction Results

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

<table>
<thead>
<tr>
<th>METHOD / FEATURES</th>
<th>BuiltEnv</th>
<th>Social [total]</th>
<th>Social [type]</th>
<th>BuiltEnv+Social [total]</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.59/0.31</td>
<td>0.52/0.27</td>
<td>0.54/0.31</td>
<td>0.58/0.34</td>
</tr>
<tr>
<td>Decision Trees (DT)</td>
<td>0.46/0.34</td>
<td>0.48/0.26</td>
<td>0.49/0.30</td>
<td>0.56/0.31</td>
</tr>
<tr>
<td>Bagging DT (BAG)</td>
<td>0.60/0.43</td>
<td>0.52/0.29</td>
<td>0.57/0.40</td>
<td>0.62/0.36</td>
</tr>
<tr>
<td>Random Forest (RF)</td>
<td><strong>0.62/0.45</strong></td>
<td>0.54/0.30</td>
<td>0.57/0.39</td>
<td>0.63/0.37</td>
</tr>
<tr>
<td>Gradient Boosting (GBoost)</td>
<td>0.60/0.41</td>
<td>0.55/0.31</td>
<td>0.58/0.41</td>
<td>0.62/0.40</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.57/0.37</td>
<td>0.55/0.34</td>
<td><strong>0.59/0.43</strong></td>
<td>0.62/0.37</td>
</tr>
<tr>
<td>Baseline</td>
<td>0.45/0.13</td>
<td>0.45/0.13</td>
<td>0.45/0.13</td>
<td>0.45/0.13</td>
</tr>
</tbody>
</table>
Prediction Results

<table>
<thead>
<tr>
<th>METHOD / FEATURES</th>
<th>Social [total]</th>
<th>BuiltEnv+Social [total]</th>
<th>BuiltEnv+Social [type]</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td>0.52/0.27</td>
<td>0.58/0.34</td>
<td>0.58/0.36</td>
</tr>
<tr>
<td>Decision Trees (DT)</td>
<td>0.48/0.26</td>
<td>0.56/0.31</td>
<td>0.52/0.36</td>
</tr>
<tr>
<td>Bagging DT (BAG)</td>
<td>0.52/0.29</td>
<td>0.62/0.36</td>
<td><strong>0.65/0.42</strong></td>
</tr>
<tr>
<td>Random Forest (RF)</td>
<td>0.54/0.30</td>
<td>0.63/0.37</td>
<td>0.63/0.41</td>
</tr>
<tr>
<td>Gradient Boosting (GBoost)</td>
<td>0.55/0.31</td>
<td>0.62/0.40</td>
<td>0.64/0.44</td>
</tr>
<tr>
<td>XGBoost</td>
<td>0.55/0.34</td>
<td>0.62/0.37</td>
<td><strong>0.65/0.44</strong></td>
</tr>
<tr>
<td>Baseline</td>
<td>0.45/0.13</td>
<td>0.45/0.13</td>
<td>0.45/0.13</td>
</tr>
</tbody>
</table>
Fig. 7. Confusion matrix and color-coded map showing the differences between ground truth and predicted safety levels for the street segments in Washington, D.C.. In (a), darker colors represent higher accuracy. In (b), green colors represent small, one level differences in the prediction, while the red color represents the worst prediction results (largest level difference).

4.3.1 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 3.2, 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 3. 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 3 shows, oversampling slightly improved the XGBoost classifier by 1% (with $M_F1 = 0.64$, $M_F1 = 0.44$) when no feature selection and a regular SMOTE were used over both built-in environment and social features (by type); but it did not improve the second best classifier, Bagging (with mRMR and SVM-SMOTE), which maintained its accuracy at $M_F1 = 0.65$, $M_F1 = 0.42$. Undersampling did not improve any F1 score.

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 3 (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 ($M_F1 = 0.78$, $M_F1 = 0.33$).
Improving Predictions

• Imbalanced dataset
  – Over/under-sampling with SMOTE
  – XGBoost only improved 1%

• Spatial Autocorrelation with Moran’s I
  – Enhance feature vector with spatially autocorrelated features from nearby segments (<150m)
  – Improved macro F1 scores by 4%
Improving Predictions

- Weighting safety labels by Familiarity and Cycling Experience boosts 1%-3%
  - Familiarity/not
  - Cycling Experience: fearless, confident, interested, reluctant
Improving Predictions

- Three (0.88/0.60) or Four (0.70/0.51) classes improve results and macro values

<table>
<thead>
<tr>
<th>METHOD</th>
<th>micro/Macro-F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Five-class (XGBoost, I&gt;0.68)</td>
<td>0.66/0.48</td>
</tr>
<tr>
<td>Four-class (GBoost, I&gt;0)</td>
<td>0.70/0.51</td>
</tr>
<tr>
<td>Three-class (XGBoost, I&gt;0)</td>
<td><strong>0.88/0.60</strong></td>
</tr>
</tbody>
</table>
Important Predictive Attributes

• XGBoost:
  – Closeness centrality of the segment,
  – Presence of cycling facilities,
  – Crime rates, and
  – Slope
Predicted Map
Future Work

• Safety perceptions and route choice
  – Combine safety predictions with data from micro-mobility solutions

• Understand changes in safety perceptions due to interventions

• Safe cycling accessibility across communities
Thank You!

Vanessa Frias-Martinez
vfrias@umd.edu