Identification of Bicycle Demand from Online Routing Requests

Hartwig H. HOCHMAIR

Abstract

Governments at all levels aim to increase cycling and walking within the mix of transportation modes. Accurate estimation of existing and potential bicycle trips on different parts of a road network is necessary to determine for which segments an investment in improved bicycling infrastructure is most effective. This paper introduces the novel idea to estimate bicycle demand on road segments based on logged trip requests that users submitted to a Web based bicycle trip planner. As a first step in this research direction, this paper assesses the general suitability of logged trip data for modeling cycling demand. More specifically, this study analyzes logged trip origins and destinations from user requests collected over a one-year period. The requests were submitted to an online bicycle trip planner developed for Broward County, Florida. The study then compares (a) point positions of logged trip origins with origins of bicycle commute trips obtained from census data, and (b) trip length distributions of logged trips with trip lengths obtained from observed bicycling trips in street networks. Several basic spatial and temporal filters are introduced and applied on the logged trip data to identify requests that most likely represent an actual trip and therefore provide a potential resource to predict bicycle demand.

1 Introduction

In many metropolitan areas, the increase of motorized traffic has contributed to societal, economic, and environmental problems. These problems have raised the interest of governments at all levels to encourage non-motorized modes of travel, including walking and cycling. Governmental efforts in improving bicycle infrastructure result in growing networks of interconnected bicycle paths, an increasing number of bicycle sharing systems, and improvements of bicycle parking facilities around transit stations to facilitate intermodal transportation (HAGELIN 2007). For future planning of bicycle infrastructure it is important to know where an investment in improved bicycle facilities would lead to benefits for cyclists and to an increase in cyclist volume.

Forecasting pedestrian and cycling travel demand for existing and proposed transportation facilities is still a big challenge for transportation planners. Traditionally, modeling bicycle demand is derived to some extent from standard forecasting models for car travel, which has limitations (KRIZEK et al. 2009). Pros and cons of alternative prediction methods, such as comparison studies, aggregative behavior models, or sketch planning, can be found in the transportation literature (KRIZEK et al. 2006). Whereas socioeconomic and environmental factors that encourage cycling are well explored (TILAHUN et al. 2007, HARVEY et al. 2008) consistent prediction of bicycle demand over a larger area is difficult (IACONO et al. 2010). LANDIS & TOOLE (1996) developed a GIS based Latent Demand Score (LDS) model which estimates the potential demand for bicycle travel that would occur if a bicycle
facility existed on a road segment. In a related effort, eliminatory constraints based on minimum acceptable Level of Service and maximum acceptable detour were included in a Latent demand model to predict the increase in bicycle demand through improved street conditions (HOCHMAIR 2009).

While most of the previously mentioned models rely on information related to trip generators, trip attractors, and road condition, Internet based data sources are at current rarely used for modeling travel trajectories and travel demand. Some related studies analyze geographic footprints of shared digital photographs uploaded onto the Internet by individual users to analyze people’s travel paths through a city, favorite places, or scenic routes (GIRARDIN et al. 2008, SCHLIEDER & MATYAS 2009, HOCHMAIR 2010). This paper proposes a novel approach to derive bicycle demand through analyzing log data collected on a Web based bicycle route planner. The basic idea is to utilize from each Web based trip request the geographic location of trip origin and destination to better understand where cyclists travel or plan to travel.

The remainder of the paper is structured as follows: Section 2 provides an overview of the modeling approach, which is followed by the analysis of the logged trip requests in section 3. Section 4 provides a summary and directions for future work.

2 Modeling Approach

This section describes the various data sources, filtering procedures for logged online data, and evaluation criteria of different filter combinations based on a comparison of spatial trip patterns between online trip requests and other reference sources. The term model refers to a given combination of filters that are applied simultaneously to the logged trip data. The filters remove some of the logged data before their comparison to other reference data.

2.1 Log Data

The University of Florida developed in cooperation with Florida International University a Web based bicycle route planner (http://bikebroward.fiu.edu/mpobike/) for Broward County, Florida. It allows the user to choose between five optimization criteria (fast, safe, simple, scenic, short), to set intermediate waypoints between trip origin and destination, and to opt for a return route from the last waypoint entered, which gives a round trip. Whenever the user requests a trip, search information is logged on the server. This includes client IP address, client hostname, geographic latitude and longitude of waypoints, time stamp, and chosen route type. After obtained the computed route, the user has the option to participate in a short survey on trip related questions, such as whether he or she plans to cycle that route, or whether the trip request was just submitted for testing purposes.

For this paper log data between December 2009 and January 2011 are analyzed. The raw data for this period contain 2517 trip requests. This includes repeated trips for a given origin-destination pair where only a change in route selection criteria was requested by a user. Removing these duplicate trip requests together with the removal of declared “test” requests from the subsequent voluntary user survey resulted in 1407 remaining trips with different origin-destination combinations. A major challenge for the prediction of bicycle demand from online data is to separate between requested trips that represent actual trips,
and those that are being made to explore the functionality of the Web application. The goal of the next steps is therefore to determine a filter method for log entries that provides Web based trip data which closely reflect actual bicycle demand. We introduce three log filter methods (models), which could be refined and improved in the future.

2.2 Model Evaluation Criteria

At current no field count data on bicycle trips exist for Broward county that could be used to compare actual trip origins, trip destinations, etc., with corresponding trip characteristics obtained from online bicycle trip requests. Therefore, we need to rely on alternative data sources for comparison and evaluation of derived trip characteristics instead. For the comparison of trip origins, we use census 2000 data from the decennial census collected by the U.S. Census Bureau (www.census.gov). The data contain the number of bicycle commute trips that originate from individual census block group polygons. This information is compiled from the census long-form, which needs to be completed by about 16% of all households. An alternative source would be data from the American Community Survey (ACS), which is an annual survey of approximately 2.5% of households, replacing the long-form portion of the decennial census.

Besides trip origins, we evaluate filter techniques regarding the comparison of average trip length and trip length distribution between online trip requests and other reference data. Reference data here relate to cycling information obtained from census data, surveys, or trips actually observed on a road network. If the average trip length from online route requests is similar to that from the reference data, this is indicative of a suitable filter method that helps to remove trips that were requested for testing the route planner only. However, the average trip length alone is inadequate to determine the model performance. This is because two different spatial travel patterns, in terms of location of trip origins and destinations, could exhibit similar average trip lengths. Therefore, a coincidence ratio \( c \) was calculated and used to compare the trip length distributions between reference data and logged trips (after applying various filter combinations). A coincidence ratio (see (Eq. 1) ranges between zero and one, with zero indicating two disjoint distributions and one indicating two identical distributions (ZHAO et al. 2004).

\[
c = \frac{\sum_i \min \left( \frac{f^m(i)}{F^m}, \frac{f^0(i)}{F^0} \right)}{\sum_i \max \left( \frac{f^m(i)}{F^m}, \frac{f^0(i)}{F^0} \right)}
\]

(Eq. 1)

where

- \( f^m(t) \) = frequency of trips in distance interval \( i \) from model;
- \( f^0(t) \) = frequency of trips in distance interval \( i \) from reference data;
- \( F^m(t) \) = total trips distributed from model;
- \( F^0(t) \) = total trips distributed from reference data;
- \( N \) = number of distance intervals.
2.3 Log Filtering Methods

Since a Web user is usually not willing to take the extra voluntary step to indicate whether he or she will actually travel the route recommended by the bicycle route planner, some patterns in the log data can be a potential source to make this distinction instead. First, one can analyze the distribution of trip requests among all users. Specific user groups of the route planner (e.g., the developer team) will frequently test the application and make simulated routing requests. This would lead to a spatial distortion of the trip origin pattern when included in the analyzed data. Therefore we suggest in a first step to remove records associated with client IP addresses that show an unusually high number of trip requests. The graph in Fig. 1 plots the user ID, which is based on a Web client’s IP address, against the number of trip requests during that one-year period. Visual inspection of the graph indicates a comparatively high number of trip requests for three users, which suggests that they are involved with application development or testing. In consequence, trip requests from these three users were removed from further analysis when applying this filter.

![Graph showing number of trip requests](image)

**Fig. 1:** Number of trip requests from different client IP addresses (total=1407)

Next, a temporal filter is suggested. Because of promotional activities of the Web site during organized events and workshops (e.g. in schools or agencies) there appear to be several days with a higher than usual number of requests (Fig. 2). There is the possibility that not all participating users would actually cycle along the requested routes, but that at least some of them would just be testing the application instead. The figure below shows six apparent outliers with more than 20 requests. Requests from these six days were therefore removed when applying this filter.

The third filter applies to geographic location. Although routing requests can be made from all over the world, it is more likely that users sending a routing request from somewhere in Broward county or even Florida actually make a planned trip rather than Web users from out of state that may just test the route planner. The logged client IP address can be paired to a geographic location. Free geolocation services are accurate between 50% and 75% for IP-to-region (or city), with improved accuracy at the state level. For this analysis, the state
of Florida was used as the geographic reference area. Thus, the third filter, if applied, would only keep trips requested from client IP addresses that can be mapped to a location in Florida.

![Graph showing the number of daily trip requests](image)

**Fig. 2:** Number of daily trip requests (total=1407)

### 3 Model Evaluation

For the evaluation part we applied all possible combinations of the three filters on the logged data, which gives a total of eight combinations (models). The trips remaining after the filtering process were then evaluated regarding the three criteria introduced in section 2.2.

#### 3.1 Location of Trip Origins

For this evaluation, locations of Web based trip origins were compared to bicycle counts in census block group polygons. We used an approach that is usually applied to quantify land cover classification accuracy using an error matrix. More specifically for each filter combination a two-by-two contingency table was constructed. The first cell counted census polygons that had a positive census bike count value and also at least one origin point from the Web based trip requests within its boundaries (positive-positive case). The next cell counted polygons with a positive census bike count and no Web based origin points within (positive-negative), and so on. From these matrices the overall accuracy for each filter combination was computed as the sum of main diagonal counts (correct matches) over the total number of polygons in the area (total=690 polygons). As can be seen from Table 1 (left two columns) without the Florida filter the model performs sometimes worse than random, whereas models limiting requests to Florida show an accuracy of over 50% (right two columns). Best results are retained if all three filters are applied simultaneously. Whereas this accuracy number is low it must be noted that the census data refer only to
commuter counts, whereas a good portion of the Web requests may have been planned for recreational trips. Some corrections for this fact would need to be included for further model improvement as part of future work.

**Table 1:** Overall accuracies of Web based trip origin classification (in %)

<table>
<thead>
<tr>
<th></th>
<th>All regions</th>
<th></th>
<th>Florida</th>
</tr>
</thead>
<tbody>
<tr>
<td>Keep peak days</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td>47.5</td>
<td>48.8</td>
<td>52.6</td>
</tr>
<tr>
<td>n</td>
<td>49.0</td>
<td>50.4</td>
<td>55.1</td>
</tr>
<tr>
<td>Keep frequent IP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>y</td>
<td></td>
<td>49.0</td>
<td>55.1</td>
</tr>
<tr>
<td>n</td>
<td></td>
<td>50.4</td>
<td>56.8</td>
</tr>
</tbody>
</table>

In order to visualize a spatial pattern of mismatches between Web based and census data, two maps based on Kernel Density Estimation (KDE) were generated. The first map takes the location of trip origins from the Web survey with all three filters applied. The second map is based on census block group centroids with associated bicycle commute trips as weights. The census bike counts in polygons (total=3406 counts) were first corrected to match the total number of Web based trip origins (total=580 features). This was done to get a comparable value range for both KDE maps. Next, the difference between KDE values from both maps was computed, where the census based KDE value was subtracted from the Web based KDE value. Fig. 3 shows the urbanized part of Broward County with the KDE difference as the background layer. Round dots indicate trip origins from Web requests. Filled polygons indicate census block groups polygons with a positive number of reported bicycle commuter trips whereas the other polygons are set to transparent. Since the brightness value of the computed difference map ranges from black to white, intermediate grey values show a good match between both data sources.

Brighter spots (more towards white) indicate locations of higher relative density within the online trip requests compared to relative densities within census based trips. Some possible explanations for differences are as follows:

- New residential housing: Whereas online routing requests may have been submitted from new residential housing areas, these areas are not yet reflected in the census 2000 data. An example is the north west region of the map.
- Recreational sites: Some Web users choose their trip origins at a recreational site which is not covered in the commuter based census data. An example is Markham park in the western region of the map north of I-595.
- Repeated client IP address: This indicates frequent requests by a single user, with the counts still being below the removal threshold of the first filter. An example is the bright spot in the map center where all requests were submitted by one user.

Darker spots indicate the reverse relationship. Possible explanations for this phenomenon are as follows:

- Residents in low household income areas often times need to rely on alternative transportation modes (e.g., cycling, walking, transit) due to reduced car ownership, while at the same time having less access to the Internet. Such locations can, for example, be found in the north and south regions along I-95.
- Areas of short commute distance: These are areas where no trip planner is required by the residents. This could explain the dark spot in the central business district of Fort Lauderdale in the center-right area of the map.

![Difference between Kernel density estimations](image)

**Fig. 3:** Difference between Kernel density estimations

### 3.2 Average Trip Length and Trip Length Distribution

Travel impedance reduces the amount of trips between two areas. For cycling, trip distance is one of the main factors of travel impedance, which means that shorter trips are more frequently made than longer ones, as illustrated in Fig. 4a for observed one-way bicycle commute trips. Impedances vary between different trip purposes. For example, people are typically willing to cycle a greater distance to work than they are to pick up a convenience item at a neighborhood store. This phenomenon is reflected in national survey data, as depicted for three trip purposes in Fig. 4b. The transportation literature suggests average cycling commuter distances between 3.3 and 5.5 km (AULTMAN-HALL et al. 1997). Although some of the requested Web trips used in this study may have been planned for recreational trips we use the cumulative frequency curve in Fig. 4a as reference distribution curve. Further we take the reported 3.7 km as reference for average trip distance. To compute lengths of trips that were requested from the Web, the shortest path between trip origin and destination was used, independent of the criterion the user actually selected.
For the computation of average trip length, only one-way trips were considered from the Web requests to match the one-way nature of reference trips in Fig. 4a. Table 2 shows the evaluation results for the trip length criterion. All models result in a significantly larger average trip length than the reference value of 3.7 km. Part of this is caused by some outliers (some trips were over 80 km), but the pattern is also indicative of trip requests that were just test runs. As opposed to the trip origin accuracy criterion (Table 1), no filter combination strongly affects average trip lengths. Similarly, variations in the coincidence ratios are moderate with no clear optimal filter setting (Table 3). Therefore, in terms of choice between the eight filter combinations, the first evaluation provides some guidance (Table 1), whereas results from Table 2 and Table 3 are not particularly helpful.

Table 2: Average trip lengths from online routing requests (in km) for various filter settings

<table>
<thead>
<tr>
<th>Keep Peak days</th>
<th>All regions</th>
<th>Florida</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Keep frequent IP</td>
<td>y</td>
<td>16.29</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>16.74</td>
</tr>
</tbody>
</table>

Table 3: Coincidence ratios from online routing requests for various filter settings

<table>
<thead>
<tr>
<th>Keep frequent IP</th>
<th>All regions</th>
<th>Florida</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>y</td>
<td>n</td>
</tr>
<tr>
<td>Peak days</td>
<td>0.723</td>
<td>0.762</td>
</tr>
<tr>
<td></td>
<td>n</td>
<td>0.761</td>
</tr>
</tbody>
</table>

Fig. 5a and Fig. 5b show for two filter settings (underlined in Table 2) the distribution of computed trip lengths from Web requests. The gray line shows the curve from within Florida (shorter average distance) compared to the curve based on requests from all regions.
In summary it can be stated that appropriate log filters can help to receive trip characteristics from Web based cycling requests that are closer to reference data sets. All three filters combined gave the best result for the origin matching criterion, but this was not the case with the other two criteria. All three evaluation criteria indicate that more refined filters are necessary to gain a better predictability of trips.

4 Conclusions and Future Work

The results presented in this paper are a snapshot of currently available log data collected over one year. The paper introduces some filter methods of logged trip request data that remove exploratory test requests in order to increase the spatial match between Web based bicycle trip patterns and reference trip data. Since the amount of available trip request data captured during the one-year period is limited, the utilized filter techniques introduced in this paper were kept simple, which leaves ample space for further refinement of the proposed methods in the future. Since it is necessary to distinguish between actual routes and test routes, this endeavor is the focus of future work. One possibility to better understand the differences between characteristics of actual trips and test trips, is to elicit a mandatory user response in the Web interface on whether the user would make a trip between the specified origin and destination, or whether this was only a test request. Further, additional sources of reference data, such as American Community Survey data, should be used for comparison and model evaluation, as long as bicycle count data from field surveys are not available. A further possible extension necessary for predicting how often a particular trip will be taken (e.g., daily or only sporadically) is to develop a mechanism for the automated extraction of the trip purpose, such as home to school, home to work, or home to recreational site. This could possibly be determined through the land uses found around the trip origin and the destination. The trip purpose affects the frequency of a trip (e.g. number of trips per month), and the route chosen to the destination (e.g. shortest or scenic). It is therefore of relevance for the identification of highly utilized network segments.
References


