# A GPS-based Bicycle Route Choice Model for San Francisco, California*

Jeffrey Hood[1,2], Elizabeth Sall[1], and Billy Charlton[1]

[1] San Francisco County Transportation Authority
100 Van Ness Ave, 26th Floor, San Francisco, CA 94102

[2] Department of City and Regional Planning, University of California, Berkeley
228 Wurster Hall #1850, Berkeley, CA 94720

Tel: (415) 522-4800, Fax: (415) 522-4829
E-mail: {jeff.hood, elizabeth.sall, billy.charlton}@sfcta.org

## 1 Introduction

In recognition of the environmental and health benefits of cycling, cities around the world are promoting use of the bicycle for everyday transportation. Because of the deterrent of traffic hazards, more cycling will not be possible in auto-dominated countries such as the United States without a traffic system that is more responsive to the needs of cyclists. Creating that system in a constrained street network with limited resources will require reliable information about the trade-offs cyclists make in choosing their routes.

For example, there is an ongoing debate about the relative merits of bicycle lanes versus wide curb lanes. While there is little evidence that either is safer than the other (see e.g. Hunter et al. 1999), proponents of bicycle lanes predict that the perception of safety will attract new cyclists (Wilkinson et al. 1994), and detractors worry about motorists' perceiving bicyclists as illegitimate users of ordinary roads (Forester 1994). Route choice models can provide the information needed to settle this debate. If inexperienced or infrequent cyclists are willing to travel farther out of the way to use bicycle lanes than experienced or frequent cyclists, a lane striping program would be expected to succeed. Otherwise, a denser network of signed, cyclist-priority, shared lanes may be a preferable and more attainable alternative.

In addition to informing design, route choice models are needed to improve demand forecasting. While cycling is increasingly being incorporated into models of mode choice, its utility function is typically represented by gross approximations of alternative-generic variables that are more predictive of automotive travel. Feeding the logsums from an estimated route choice model back into the mode choice utility specification would provide a measure of cycling-specific accessibility with which one could test the effect of network improvements on cycling mode choice. Route choice predictions also turn distributed trips into the link-level volume assignments necessary to target operational improvements where they are needed most.

*This research is supported by a Planning and Research Grant from the California Department of Transportation.

Until recently, information about cyclist route choice came almost exclusively from stated preference surveys because the cost of data collection and computational complexity of high-resolution network algorithms limited revealed preference studies to small samples and descriptive analysis (e.g. Aultman-Hall et al. 1997). A recent stated preference study (Sener et al. 2009) contains a thorough review of the existing research as of 2007; none of 16 the studies listed applied multivariate analysis to revealed preference data. While good experimental design can improve the external validity of stated preference research, the results must still be verified by actual behavior (Bradley 1988).

Revealed preference route choice research is no longer impractical. The radical cost reduction and widespread adoption of Global Positioning System (GPS) devices have made possible the collection of route data on a previously unimaginable scale. A recent technical report demonstrated the results of the first bicycle route choice model based on a large sample of GPS observations in the city of Zürich, using trips extracted from a larger long-duration GPS dataset using a mode detection algorithm (Menghini et al. 2009). The study found that cyclists tolerate only short detours from the minimum distance path. However, the set of additional variables analyzed—number of traffic lights, terrain gradient, and utilization of bicycle facilities—was small.

This paper presents the results of the first GPS-based bicycle route choice model for the United States, developed at the San Francisco County Transportation Authority (SFCTA) as part of SF-CHAMP, the San Francisco Chained Activity Model Process (Outwater & Charlton 2006). It adds to the findings in the Zürich study in several ways. First, it updates the results for the US context, where street networks, bicycle facilities, and behavior may differ. Second, rather than extracting routes from a raw GPS dataset without accompanying information, the study takes advantage of the large user base of GPS-enabled smartphones, collecting data through a free application called CycleTracks, available at `http://www.sfcta.org/CycleTracks`. Apart from ease of distribution and increased sampling of the small population of cyclists, the advantage of this approach is the ability to record personal characteristics and trip purpose, which is especially important for bicycle modeling given the extreme difference between utilitarian and recreational travel. Third, the current study uses a new method of choice set generation: a network-based automatic calibration of the "doubly stochastic" method of Bovy & Fiorenzo-Catalano (2007). This method produces more heterogeneous, attractive, unbiased, and exhaustive choice sets than other methods, yielding more reliable parameter estimates and predictions. Finally, the study analyzes a richer set of attributes of the network and environment, including different types of bicycle facility, freeflow speed, number of lanes, number of turns, weather, daylight, crime, and traffic volume, which has been shown to be the most important factor in stated preference surveys.

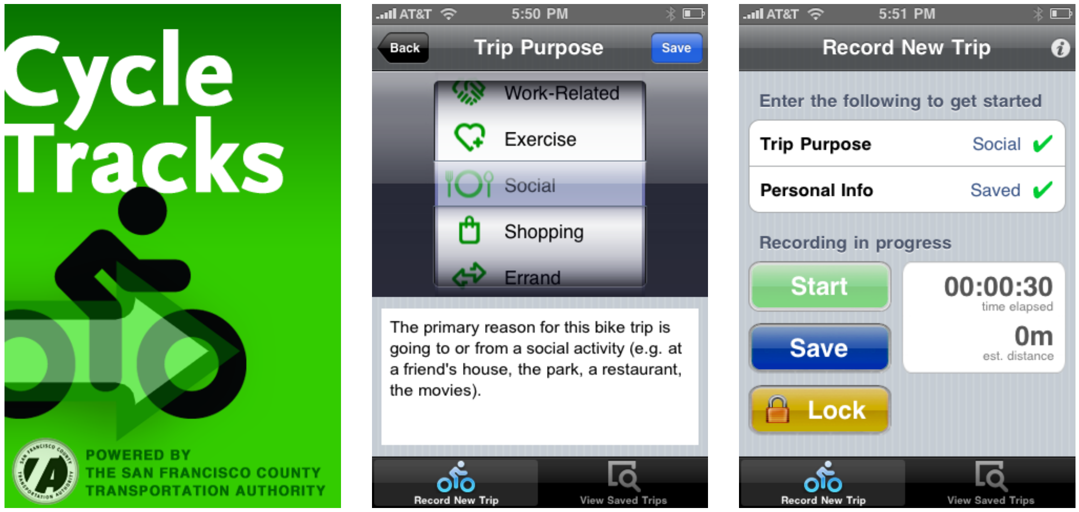



Figure 1: CycleTracks for the Apple iPhone

# 2 Methods

## 2.1 Travel Data Collection

GPS data of cyclists' routes were collected using CycleTracks, an application for the Apple iPhone and Google Android smartphone platforms that was developed for this study. The interface was slightly different for the two platforms, but generally worked as follows. At the beginning of each trip, the user selected a trip purpose and then the phone recorded GPS coordinates until the user canceled the trip or indicated that the trip was complete. At this point, the user reconfirmed the trip purpose, entered a comment to accompany the trip (if desired), and submitted the data to the web server. As one of the incentives for using the application, the user could, when not recording, view a list of saved trips with maps and simple statistics such as distance, time, and average speed. The interface for the application appears in Figure 1.

The user also had the option of entering personal information: age, sex, frequency of cycling, home and work zip code, and e-mail address. The personal information was linked to the phone rather than the trip, but we assumed the number of phones with multiple users was negligible. Each user that entered an e-mail address and uploaded at least one bicycle trip was entered in a drawing to win one of several $50 iTunes gift cards.

The application was distributed online through the iTunes App Store, the Android Market, and the SFCTA website. To recruit participants, we sent e-mail announcements and links to local bicycle coalitions and university groups, asking them to forward them to their members. We also disseminated information about the study to media outlets in the Bay Area. For more details on the GPS data collection methods, see Charlton et al. (2010). To download CycleTracks, visit http://www.sfcta.org/CycleTracks.

## 2.2 GPS Post-Processing

Between November 12, 2009, and April 18, 2010, 1,083 users downloaded the application, and 952 submitted at least one trip. Including all data in and out of the Bay Area, 7,096 trips were collected. To obtain the most detailed model specification possible, we restricted our analysis to the City of San Francisco, where we have the richest set of network attributes, and to non-exercise trips, for computational feasibility and because many exercise trips lack a true destination. After this restriction, 5,178 trips remained.

Next, we cleaned and smoothed the data, and identified intermediate destinations and changes of travel mode by analyzing idle times, speeds, and accelerations using the fuzzy logic method of Schüssler & Axhausen (2009*b*). The mode detection algorithm labeled an excessively large proportion of the GPS traces as transit trips because of high 95th percentile speeds. Therefore, we relabeled any such trace as a bike trip, as long as the algorithm found no activities and only one stage. Finally, the GPS points were allocated to the street network using the map matching algorithm of Schüssler & Axhausen (2009*a*). After processing, 3,034 bicycle stages from 2,777 traces uploaded by 366 users were successfully matched to the network.

## 2.3 Participants

Because participation was limited to smartphone users, and because the greatest selection rate likely occurred among members of the bicycle coalitions that helped promote the application, the sample is biased. However, this drawback was outweighed by the advantages of the data collection method: reduced cost, increased rates of sampling for the small population of cyclists, and the ability to record personal characteristics and trip purposes.

Of those users with data remaining after GPS processing who reported age (N=297), the mean was 34 with a standard deviation of 9. Of those who reported gender (N=292), 21% were female. Of those who reported cycling frequency (N=270), 60% bicycle daily, 34% bicycle several times per week, 7% bicycle several times per month, and none bicycle less than once a month.

To test for bias in our sample, we compared it to the sample of people in the 2000 Bay Area Travel Survey (Morpace International 2002) who reported at least one unlinked cycling trip with an origin or destination in San Francisco during the two-day activity diary. In this sample (N=153), the mean age was 33 with a standard deviation of 12, and 35% were female. A Welch's $t$-test for independent samples with different variances showed no significant difference in the mean age between the samples ($t = 1.0, df = 235, p = 0.31$). However, a $z$-test for Bernoulli proportions in independent samples did show that the lower proportion of females in the smartphone sample was significant ($z = -3.5, p = 0.00$)[1].

[1] In applying the tests, we used variance estimates for the BATS sample that ignored non-random sources of sampling error because BATS provides no jackknife replicate weights. Thus, the magnitude of the $t$- and $z$- statistics may be inflated.

We did not collect data on other characteristics for which we expected bias, such as income, because we expected a poor response rate due to privacy concerns. Nonetheless, we conjecture that sample bias is a negligble problem for route choice modeling because the choice is conditioned on already having chosen a destination and mode, and should depend less on demographics and more on characteristics of the alternatives.

## 2.4 Network and Environmental Attributes

The network model was created by integrating Geographic Information Systems data from multiple sources into the network file maintained by the SFCTA. The San Francisco portion of the network has 33,575 links and 10,234 nodes. Definitions of the hypothesized relevant attributes—length, speed limit, traffic volume, type of bicycle facility, street slope, and local crime rate–appear in Table 1.

The bicycle facilities are designated Class I (bike path), II (bike lane), or III (bike route). Bike paths are off-street facilities for the exclusive use of bicycles and pedestrians, and exist in San Francisco primarily only in parks, along coastlines, and as overpasses. Bike lanes are striped in the roadway at a width of about 5 feet, and are for the exclusive use of cyclists except when vehicles are turning or parking. Bike routes are shared with vehicle traffic, and are indicated by signs or pavement markings. San Francisco contains 23 miles of bike paths, 45 miles of bike lanes, and 132 miles of bike routes[2]. Forty percent of bike routes have an outside lane width of 14 feet or greater, and 17% have "sharrows" (San Francisco Municipal Transportation Agency 2009).

In addition to network attributes, we selected two environmental variables to describe the choice context. Hourly rain in inches came from http://www.wunderground.com/, and sunrise and sunset times came from http://www.mindspring.com/~cavu/sunset.html.

## 2.5 Choice Set Generation

Route choice modeling requires, for each origin-destination pair, the identification of a set of alternative non-chosen routes. In large networks, the universal choice set is typically of unknown size, and candiates must be extracted from the network. Since the quality of model estimates and predictions depends heavily on the size and composition of the choice sets (Prato & Bekhor 2007), several methods have been proposed, but only methods based on repeated shortest path searches have been proven in large networks. The most promising of such methods are link elimination, which was used in the Zürich bicycle route choice study as detailed in Schüssler et al. (2009), and stochastic path generation. While these methods suffice for auto route choice, which depends heavily on a single impedance attribute—travel time—they are

[2]These measurements are neither uni-directional nor bi-directional. That is, the length of one-way and two-way links are counted equally in the total.

| Attribute | Source | Description |
|---|---|---|
| Length | SFCTA | Arc length in miles |
| Traffic Volume (5 periods) | SFCTA | Directional vph from SF-CHAMP assignment |
| Freeflow Speed | SFCTA | In mph |
| Lanes | SFCTA | Directional no. of lanes |
| Bike Class | Metropolitan Trans. Commission | Type of bike facility (I-III) |
| Up-slope | City & County of San Francisco | Rise in 5-ft. contours ×100 /arc length betw. contours |
| Crime | San Francisco Police Department | Annual no. violent crimes per sq. mi. within 1/10 mi. |

Table 1: Network Attributes

unsatisfactory for bicycle route choice, which depends on a variety of environmental variables.

Building on simple stochastic path generation, Bovy & Fiorenzo-Catalano (2007) present a “doubly stochastic” method in which both link attributes and generalized cost coefficients are randomized for each shortest path search. This method produces the chosen route for a high proportion of observations, but the prior distribution from which the coefficients are drawn must be calibrated. The authors suggest either (1) using a distribution with a mean equal to values already found in the literature, which is not possible for novel model estimations and may lead to bias in any case, or (2) optimizing the distribution based on the proportion of observations reproduced, which may lead to endogeneity.

We eliminated this need for calibration and danger of bias or endogeneity by extracting the prior distribution of coefficients from the network. Given an origin-destination pair $(s, t)$ and a vector of link attribute functions $x : A \to \mathbb{R}^n_{\geq 0}$, where $A$ is the set of links in the network, define the attributes of a path $\Gamma$ as $x(\Gamma) = \sum_{a \in \Gamma} x(a)$, and consider the set of all possible path attribute vectors $X = \{x(\Gamma) \mid \Gamma \in \mathcal{U}_{st}\}$, where $\mathcal{U}_{st}$ is the set of all paths from $s$ to $t$. For every vector of coefficients $\beta$ and scalar $r$, consider the half-spaces $Y_{\beta,r} = \{y \mid \langle \beta, y \rangle \geq r\}$, where $\langle \cdot, \cdot \rangle$ is the inner product on $\mathbb{R}^n$, and let $\overline{\text{conv}}(X) = \bigcap_{\beta,r :\, X \subset Y_{\beta,r}} Y_{\beta,r}$ be the upper convex hull of $X$. The boundary $\partial\overline{\text{conv}}(X)$ of $\overline{\text{conv}}(X)$ is the attribute production possibility frontier of the network. Our goal is to sample paths with attributes near $\partial\overline{\text{conv}}(X)$ with approximately uniform probabilities.

The choice set generation algorithm proceeds in two phases. First, the distribution of coefficients that produces the attribute possibiltiy frontier is extracted from the network in a preprocessing phase. Second, the choice sets are generated using doubly stochastic shortest path searches.

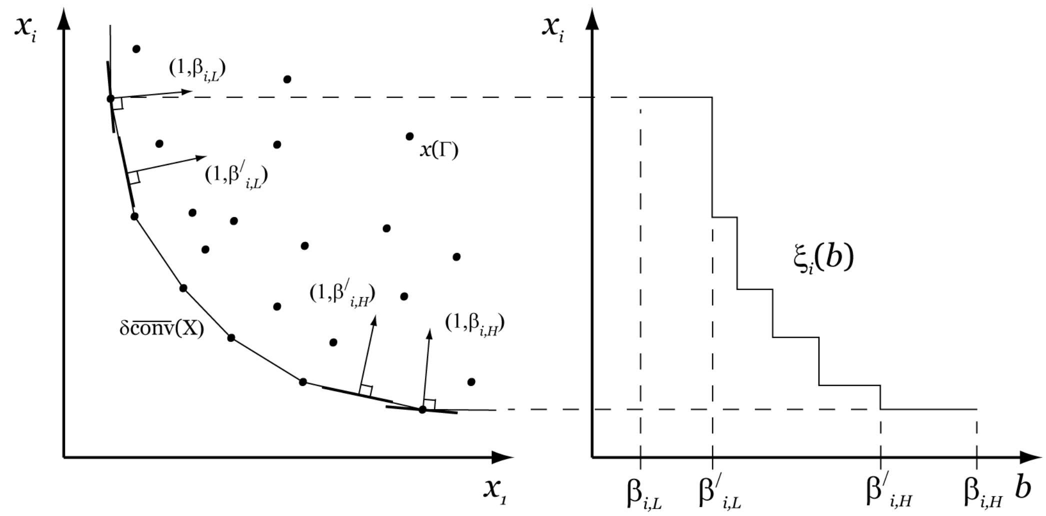



Figure 2: Extracting the prior coefficient distribution

### Preprocessing Phase: Prior Coefficient Distribution Extraction

Select as a reference variable $x_1$ one attribute with non-zero values for each link, set its coefficient $\beta_1 = 1$, and establish large boundary intervals $[\beta_{i,L}, \beta_{i,H}]$ for coefficients $i > 1$. Define the following notation for the replacement of the $i$-th coordinate of a vector $v$ by the value $\lambda$: $v(i \to \lambda) = (v_1, \ldots, v_{i-1}, \lambda, v_{i+1}, \ldots, v_n)$. For each origin-destination pair $(s,t)$, each $i$, and for any vector of coefficients $\beta$, define the function $\xi_i(b, \beta) = \min_{\Gamma \in \mathcal{U}_{st}} \sum_{a \in \Gamma} \langle \beta(i \to b), x(a) \rangle$, giving the value of attribute $i$ under generalized cost minimization with coefficients $\beta(i \to b)$. Let $e_1 = (1, 0, \ldots, 0)$ be the unit vector in direction of the reference variable. For each origin-destination pair in the observations (or a sample of observations), find new boundary values

$$\begin{array}{rcl} \beta'_{i,L} & = & \sup\{b \geq \beta_{i,L} \mid \xi_i(b) = \xi_i(\beta_{i,L}, e_1)\}, \\ \beta'_{i,H} & = & \inf\{b \leq \beta_{i,H} \mid \xi_i(b) = \xi_i(\beta_{i,H}, e_1)\} \end{array} \quad (1)$$

using logarithmic binary search. The interval $[\beta'_{i,L}, \beta'_{i,H}]$ is the smallest interval that will produce all paths with attributes on the non-trivial portion of the $x_1$-$x_i$ possibility frontier (Figure 2).

For some attributes, minimization of $x_i$ will tend to co-occur with minimization of the reference variable $x_1$, and so $\beta'_{i,L}$ and $\beta'_{i,H}$ will be ill-defined. If this occurs frequently, first calculate $\beta'_{j,L}$ and $\beta'_{j,H}$ for $j \neq i$, and perform the above with the $j$-th coordinate of the unit vector $e_1$ replaced with the geometric mean of $\beta'_{j,L}$ and $\beta'_{j,H}$ for $j \neq i$.

To obtain the prior distibtuion to be used in the generation phase, let $\beta^*_{i,H}$ be the geometric mean of the higher bounds $\beta'_{i,U}$ from the sample of observations for which a range of coefficients is found, and let $\beta^*_{i,L}$ be the geometric mean of the extracted lower bounds. The prior distribution is log-uniform on the Cartesian product of the intervals $[\beta^*_{i,L}, \beta^*_{i,H}]$.

| Param. | Description | Value |
|---|---|---|
| $M$ | No. random generalized cost coefficient samples | 32 |
| $N$ | No. link-randomized searches per coefficient sample | 3 |
| $d$ | Percent overlap filtering threshold | 90% |
| $\sigma$ | Link randomization scale parameter | 0.7 |

Table 2: Choice set generation parameters

### Generation Phase: Doubly Stochastic Shortest Path Searches

For $M$ samples of $\beta$ from the prior distribution, and each origin-destination pair $(s,t)$, search for the path $\Gamma \in \mathcal{U}_{st}$ that minimizes the generalized cost $\sum_{a\in\Gamma}\langle\beta, x(a)\rangle Z_a$, for $N$ different samples of link randomization values $Z_a \sim \text{Unif}(1-\sigma, 1+\sigma)$.[3] Add the chosen path, and filter the generated set until no paths have overlapping lengths of over $d$%.

### Parameter Selection and Implementation

The choice set generation method was implemented in Python, as an extension of the NetworkX package (Hagberg et al. 2008). The parameters subject to analyst judgment and calibration appear in Table 2.

At first, all attributes that we tested in model estimation were used in choice set generation. Subsequently, attributes that were not statistically significant were eliminated from choice set generation to improve sampling rates, and the model was re-estimated. However, we found it necessary to include one more attribute in the choice set generation than was used in the utility specification. Without this additional dimension (we selected length $\times$ daily traffic), a correction for route overlap (see section 3.2) could not be introduced into the utility function without near multicollinearity of attributes within many choice sets. The attributes used for choice set generation in the final version of the model appear, along with the initial and extracted prior distributions for the generalized cost coefficients, in Table 3.

### Properties of the Choice Sets

To evaluate the quality of the choice sets, we compared their properties to those generated using link elimination as in Schüssler et al. (2009). Choice sets were generated for all observations except for a random sample of 10% of the observations that were held back from estimation. The average number of unique routes in the doubly stochastic choice sets was 76, and 51 after filtering. The link elimination choice sets contained 96 routes each. Despite using the faster A* search algorithm, link elimination is not faster than the doubly stochastic method which uses Dijkstra because it frequently finds the same route again and again. Including the time to extract the

[3]Here our method differs slightly from Bovy & Fiorenzo-Catalano, where each link attribute value was randomized independently. We found applying link randomization after the generalized cost calculation to perform equivalently.

| Attribute | Initial Interval $\beta_{i,L}$ | $\beta_{i,H}$ | Final Interval $\beta^*_{i,L}$ | $\beta^*_{i,H}$ |
|---|---|---|---|---|
| Length (reference) | 1.0 | 1.0 | 1.0 | 1.0 |
| Length off bike paths | $1.0\times10^{-7}$ | 1,000. | 0.082 | 6.41 |
| Length off bike lanes | $1.0\times10^{-7}$ | 1,000. | 0.075 | 5.29 |
| Length off bike routes | $1.0\times10^{-7}$ | 1,000. | 0.098 | 10.8 |
| Length × up-slope (ft/100 ft) | $1.0\times10^{-7}$ | 1,000. | 0.043 | 2.36 |
| Length wrong way | $1.0\times10^{-7}$ | 1,000. | 0.031 | 0.49 |
| Number of turns* | $1.0\times10^{-7}$ | 1,000. | 0.009 | 1.63 |
| Length × daily traffic (1,000s) | $1.0\times10^{-7}$ | 1,000. | 0.010 | 9.51 |

*Extracted with other coefficients at their medians

Table 3: Extracted prior coefficient distribution

prior distribution from a sample of 500 observations, the doubly stochastic method took 4 hours 37 minutes to generate sets for 2,678 observations with 4 processors. The link elimination method took 8 hours 6 minutes.

The cumulative distribution of the choice sets' maximum overlap with the chosen route appears in Figure 3a. The ideal situation is always to produce the chosen route, so a lower area under the curve is better. Both methods produce the chosen route exactly for the easiest one-third of observations, but the doubly stochastic method vastly improves overlap for the more difficult observations.

Furthermore, the doubly stochastic routes are more similar to the chosen route in cases when the overlap is less than 100 percent. Figure 3b is a frequency distribution of the choice set's minumum of a dissimilarity index, the Normalized Euclidean Distance from the attributes of the chosen route, defined by

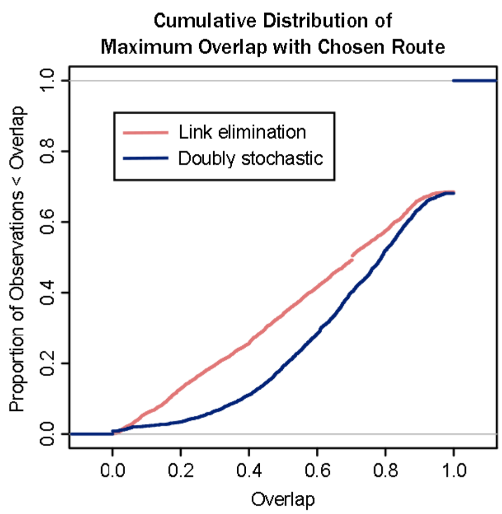
$$\mathrm{NED}_{ni} = \sqrt{\sum_{m=1}^{M} \left( \frac{x_{nim} - x^*_{nm}}{\sigma^*_m} \right)^2} \tag{2}$$

where $i$ is an alternative in the generated choice set for observation $n$, $M$ is the number of attributes in the estimated utility function, $x^*_{nm}$ is the attribute vector of the chosen route, and $\sigma^*_m$ is the standard deviation of the $m$-th attribute over all of the chosen routes. The doubly stochastic method produces routes that are similar to the chosen route much more frequently. The benefits of this similarity are less biased and more efficient parameter estimates because of the additional information gained about the decision-maker's tradeoffs.

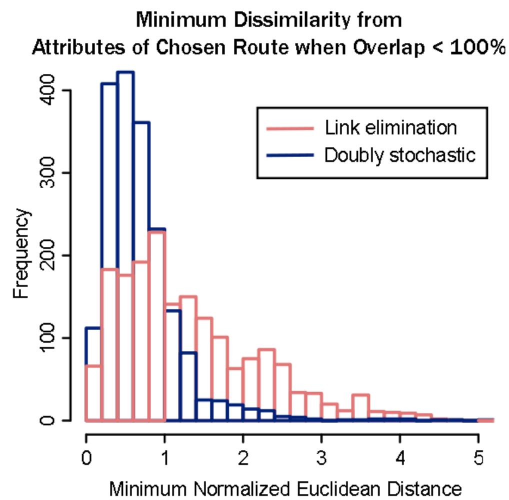
# 3 Results

## 3.1 Descriptive Statistics

As described in section 2, the GPS traces we received from users of the smartphone application CycleTracks were processed for mode and activity detection, and matched




(a) Overlap




(b) Dissimilarity

Figure 3: Choice set quality

| Attribute | Mean | Std. Dev. |
|---|---|---|
| Length (mi) | 2.23 | 1.65 |
| Turns per mile | 2.62 | 1.64 |
| Proportion wrong way | 0.02 | 0.06 |
| Proportion bike paths | 0.04 | 0.10 |
| Proportion bike lanes | 0.28 | 0.27 |
| Proportion bike routes | 0.24 | 0.23 |
| Avg. up-slope (ft/100ft) | 0.80 | 1.10 |

Table 4: Mean and standard deviation of attributes for the chosen routes

to a network model. After discarding 769 exercise trips because of computational limitations and their lack of a true destination, and restricting our analysis to San Francisco, 3,034 bicycle stages from 2,777 traces uploaded by 366 unique users were successfully matched to the network

When a GPS trace was split into multiple stages by the processing algorithm, the trace purpose was assigned to each stage. The most common trip purpose was commute (55%), followed by errand (16%), social (10%), shopping (9%), work-related (5%), other (3%), and school (1%). The means and standard deviations of the attributes for the chosen routes appear in table 4.

## 3.2 Discrete Choice Modeling Framework

The independence of irrelevant alternatives (IIA) property of the Multinomial Logit model makes it inappropriate for estimating models containing similar alternatives, where the error terms are correlated. This drawback is particularly problematic for route choice models because alternative routes actually overlap. Furthermore, incorporating the error structure explicitly is not practical because of the high number of alternatives.

The other option to correct for the IIA problem is to introduce a similarity measure in the utility function. We used the Path Size measure of Ben-Akiva & Bierlaire (1999):

$$PS_{ni} = \sum_{a \in \Gamma_i} \frac{l_a}{L_i} \left( \sum_{j \in \mathcal{C}_n} \delta_{aj} \frac{L^*_{\mathcal{C}_n}}{L_j} \right)^{-1}, \tag{3}$$

where $\Gamma_i$ is the set of links in alternative $i$, $l_a$ and $L_a$ are the length of link $a$ and path $i$, $\mathcal{C}_n$ is the choice set for decision-maker $n$, $\delta_{aj}$ is one if link $a$ is part of path $i$ and zero otherwise, and $L^*_{\mathcal{C}_n}$ is the length of the shortest path in $\mathcal{C}_n$. The corrected utility function is

$$U_{ni} = \beta \cdot x_{ni} + \beta_{PS} \log PS_{ni} + \varepsilon_{ni}, \tag{4}$$

where $x_{ni}$ is a vector of route attributes and interactions, $\beta$ is a vector of coefficients, $\beta_{PS}$ is a scalar parameter, and $\varepsilon_{ni}$ are i.i.d. Gumbel. The choice probabilities are

$$P_{ni} = \frac{\exp(\beta \cdot x_{ni} + \beta_{PS} \log PS_{ni})}{\sum_j \exp(\beta \cdot x_{ni} + \beta_{PS} \log PS_{ni})}. \tag{5}$$

The parameters $\beta, \beta_{PS}$ were estimated in BIOGEME (Bierlaire 2003) using maximum likelihood and the DONLP2 optimization algorithm (Spellucci 1988).

## 3.3 Model Estimation

The estimated parameters for the bicycle route choice model appear in Table 5. A random sample of 10% of the observations were held back to evaluate the model's prediction success on this subset separately. Because of extreme variability in the number of observations per individual in the dataset, each observation was weighted in the likelihood function by the inverse of the number of observations for the individual so that each individual would have equal weight. Interaction terms are indented in italics.

The coefficients indicate that cyclists prefer shorter routes with fewer turns, and will not go the wrong way down a one-way street without considerable savings in effort. The variables measuring the proportion of the route containing the three types of bicycle facility are all measured on the same scale. Therefore, their relative magnitudes indicate the degree to which the average cyclist prefers one over the other. The effect of bike lanes 1.89 is larger and significantly different from the mean effect of shared-lane bicycle routes (0.35) at the 5% level, indicating that bike lanes are preferred on average. The interaction of bike paths with cycling frequency indicates that infrequent cyclists tend to have a stronger preference for bike lanes, as well. The negative coefficients on the average up-slope and corresponding interactions indicate that avoidance of hills is especially strong for women and when commuting. None of the other network attributes or environmental variables could be incorporated into the model with statistically significant coefficients. The logarithm of the path size

| Attribute | Coef. | Std. err. | $t$-stat. | $p$-value |
|---|---|---|---|---|
| Length | −1.05 | 0.09 | −11.80 | 0.00 |
| Turns per mile | −0.21 | 0.02 | −12.15 | 0.00 |
| Proportion wrong way | −13.30 | 0.67 | −19.87 | 0.00 |
| Proportion bike paths | 1.89 | 0.31 | 6.17 | 0.00 |
| Proportion bike lanes | 2.15 | 0.12 | 17.69 | 0.00 |
| *Cycling freq. < several per week* | 1.85 | 0.04 | 44.94 | 0.00 |
| Proportion bike routes | 0.35 | 0.11 | 3.14 | 0.00 |
| Average up-slope (ft/100ft) | −0.50 | 0.08 | −6.35 | 0.00 |
| *Female* | −0.96 | 0.22 | −4.34 | 0.00 |
| *Commute* | −0.90 | 0.11 | −8.21 | 0.00 |
| log(path size) | 1.07 | 0.04 | 26.38 | 0.00 |

| | |
|---|---|
| Number of observations: | 2,678 |
| Null log-likelihood: | $-10,006$ |
| Final log-likelihood: | $-7,213$ |
| Adjusted rho-square: | 0.23 |

Table 5: Route choice utility function estimation results

variable corrects for route overlap. Notably, its estimated coefficient is not significantly different at the 5% level from the theoretically correct value in a model with a scale parameter of one.

## 3.4 Marginal Rates of Substitution

To avoid multicollinearity within choice sets, Length was divided into all of the other attributes in the utility specification, so direct Marginal Rates of Substitution (MRS) values are difficult to interpret, espectially the MRS for Turns per mile, which is in units of mi$^2$ per turn. Therefore we recovered the average MRS for turns, total rise, and length on links of different types by reparametrizing the attribute space and differentiating the deterministic portion of utility with respect to the new variables. The reparametrized utility function is

$$\begin{aligned} U = {} & \beta_{L_{\text{Tot}}}(L_S + L_W + L_1 + L_2 + L_3) \\ & + \frac{\beta_{\text{TpM}} T + \beta_{P_W} L_W + \beta_{P_1} L_1 + \beta_{P_2} L_2 + \beta_{P_3} L_3 + \beta_{\text{AvS}} R}{L_S + L_W + L_1 + L_2 + L_3} + C \end{aligned} \tag{6}$$

where $\beta_{L_{\text{Tot}}}$, $\beta_{\text{TpM}}$, $\beta_{P_W}$, $\beta_{P_1}$ $\beta_{P_2}$ $\beta_{P_3}$, and $\beta_{\text{AvS}}$ are the model coefficients for total length; turns per mile; proportion wrong way; proportion of Class I off-street bike paths, Class II bike lanes, and Class III shared-lane bike routes; and average up-slope, respectively;

| MRS of Length on street for | Value | Units |
|---|---|---|
| Turns | 0.11 | mi/turn |
| Length wrong way | 4.02 | None |
| Length on bike paths | 0.57 | None |
| Length on bike lanes | 0.49 | None |
| Length on bike routes | 0.92 | None |
| Total rise | 1.12 | mi/100 ft |

Table 6: Average marginal rates of substitution

$$
\begin{aligned}
L_S &= \text{Length on ordinary streets} \\
L_W &= \text{Length wrong way} \\
L_1 &= \text{Length on Class I off-street bike paths} \\
L_2 &= \text{Length on Class II bike lanes} \\
L_3 &= \text{Length on Class III shared-lane bicycle routes} \\
T &= \text{Total number of turns} \\
R &= \text{Total rise}
\end{aligned}
$$

and path size is held constant.

The marginal utilities of the length on different types of links, holding the number of turns per mile and average up-slope constant, are

$$MU_{L_S} = \beta_{L_{\mathrm{Tot}}} - \sum_{i \in \{W,1,2,3\}} \beta_{P_i} L_i / L_{\mathrm{Tot}}^2 \tag{7a}$$

$$MU_{L_j} = \beta_{L_{\mathrm{Tot}}} + \sum_{i \in \{W,1,2,3\}} (\beta_{P_j} - \beta_{P_i}) L_i / L_{\mathrm{Tot}}^2 \tag{7b}$$

and the marginal utilities of turns and rise in elevation, holding the lengths constant, are

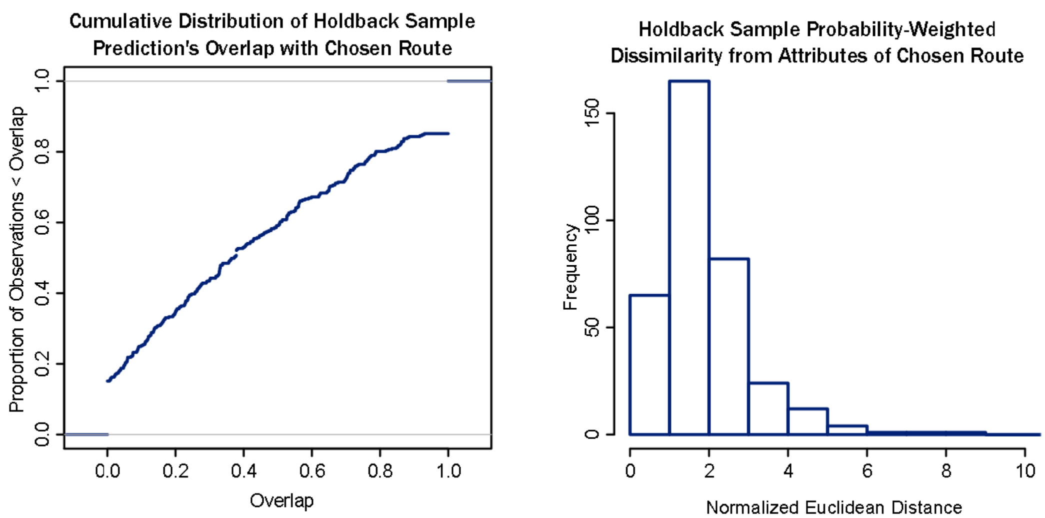
$$MU_T = \beta_{\mathrm{TpM}} / L_{\mathrm{Tot}} \tag{8a}$$

$$MU_R = \beta_{\mathrm{AvS}} / L_{\mathrm{Tot}} \tag{8b}$$

The average marginal rates of substitution appear in Table 6. The average cyclist will avoid a turn if it costs no more than one-tenth of a mile, and will avoid climbing a hill 100 feet tall as long as the detour is less than roughly one mile. The MRS for the lengths are dimensionless, and so represent the relative disutility of traversing the different types of links. Cyclists will not travel the wrong way down a one-way street unless doing so saves more than four times the distance (or its equivalent in turns or hill climbing) elsewhere. On the other hand, the average cyclist is willing to add a mile on bike lanes in exchange for only half a mile on ordinary roads.

| Attribute | $\beta_{\text{low}}$ | $\beta_{\text{high}}$ |
|---|---|---|
| Length | 1.05 | 1.05 |
| Length off bike paths | 0.42 | 1.70 |
| Length off bike lanes | 0.50 | 1.98 |
| Length off bike routes | 0.08 | 0.31 |
| Length $\times$ up-slope (ft/100 ft) | 0.35 | 1.38 |
| Length wrong way | 2.98 | 11.93 |
| Number of turns | 0.05 | 0.19 |

Table 7: Calibrated prior coefficient distribution




(a) Overlap (b) Dissimilarity

Figure 4: Prediction quality

## 3.5 Model Validation

After estimating the model, we calibrated the variance of the choice set parameters and validated the model by looking at predictions for a holdback sample of 303 observations. The calibrated link randomization scale parameter was $\sigma = 0.4$, and the calibrated lower and upper bounds for the random generalized cost coefficient distribution appear in Table 7. The advantage of reducing the variance of the coefficient distribution is that we could reduce the number of coefficient samples to 20.

The cumulative distribution of the prediction's overlap with the chosen route when the chosen route was not included in the choice set appears in Figure 4a. Fifteen percent of the observations were predicted exactly. We are not aware of any other such route choice model validations, so it is difficult to evaluate the rest of the scale without a basis for comparison. Therefore, we also looked at the predictions' probability-weighted dissimilarity from the attributes of the chosen route (Figure 4b), which was usually low.

# 4 Discussion

The path size logit route choice model estimated in this study indicates that cyclists in San Francisco strongly prefer bike lanes to other types of bicycle facility, and disfavor climbing hills, turning, and deviating excessively from the minimum distance paths. The model coefficients offer insights into best practices in bicycle network design and a measure of the user benefits of bicycle facilities, and incorporation of the route choice model into a larger demand forecasing process will greatly enhance the responsiveness of planning to the needs of cyclists and provide new abilities to test policies intended to promote increased use of this efficient travel mode.

### Comparsion to Existing Literature

The agreement of our results with those from other areas is mixed. Obviously, route length (or travel time) has generally been found to be an important factor in route selection. Another point of agreement is the influence of terrain. The revealed preference study in Zürich (Menghini et al. 2009) found that the maximum slope of a route negatively but very slightly influenced route selection and that the average slope had no effect. A mixed logit analysis of stated preferences in Texas (Sener et al. 2009) found that steep hills were disfavored more by women an commuters, which is precisely our finding.

The overall influence of cycling facilities was also consistent with these studies. While Menghini et al. found that cyclists would only go an additional 233 meters (0.14 miles) to use a continuous bicycle facility, Sener et al. found they would add 13 minutes (about 2 miles at 10 mph) to an existing commute of less than five miles. Our marginal rates of substitution, which vary by facility type, are spread between these values.

However, the strong preference for bicycle lanes in our study differed from Sener et al., where shared-lane bicycle routes were slightly preferred to bicycle lanes. It is possible that the disagreement is due to differences in the design of these facilities between the two regions; however, it is more likely due to the limitations of the stated preference approach. People often do not do as they say.

The insignificance of traffic volume in our model was surprising, given that it was the most important factor after travel time in Sener et. al. The problem was likely due to the difficulty of separating the effects of related covariates in the revealed preference appraoch, and underscores the importance of employing both stated and revealed preference methods to obtain a complete picture traveler behavior.

### Benefit-Cost Analysis of Bicycle Facilities

The marginal rates of substitution (MRS) of length on ordinary streets for length on bicycle facilities from Table 6 provide a measure of the user benefits of bicycle infrastructure that can be used in a benefit-cost analysis of bicycle facilities. For example, the MRS for length on bike lanes of 0.49 implies that, if no detour is required, the value to the user of a mile spent on a bike lane is equivalent to a savings of 0.51

miles without a bicycle lane. Assuming an average speed of 10 mph and value of time for all trip purposes of $18.82 in 2009 dollars (the average of work and non-work values used in the San Francisco model), users derive a benefit of approximately $0.96 per mile of travel on bike lanes.

Using the online cost estimation software developed as part of NCHRP Report 552, *Guidelines for Analysis of Investments in Bicycle Facilities* (Krizek et al. 2006), we estimate build year capital costs for installing a 5 foot-wide bike lane by restriping an existing road in San Francisco to be $14,073 per mile, with annual operating and maintenance costs of $6,500 per mile. Adding the opportunity cost of repurposing the land at an average value of $86 per square foot (taken from assessor's data) and amortizing the costs over 30 years, the annual cost of a new bike lane is approximately $82,649 per mile.

Therefore, adding a bike lane in San Francisco is justified *by the user benefits alone* wherever it will carry more than 235 trips per day. Considering external benefits or finding space of lower opportunity cost in excessively wide roadways would bring the benefit-cost ratio above one for much lower cyclist volumes.

### Trip Assignment

After estimating the route choice model, we added a bicycle trip assignment module to SF-CHAMP (Figure 5). The benefits of obtaining link-level bicycle volume estimates will include the identification of areas where operational improvements would benefit the greatest number of cyclists and the ability to assess the demand for new facilities, given competition with existing alternatives.

In order to produce the trip assignment in a reasonable amount of time, we projected trips originating outside of San Francisco to boundary zones by finding the first San Francisco zone encountered in the minimum distance path. Choice sets were then generated for the 981 San Francisco zones using single source searches. Rather than using different random seeds for each origin zone, 20 random seeds were drawn for each zone from a larger set of 40 random seeds to reduce computation time while avoiding coupling paths from nearby origins together. The ability to generate choice sets using single source searches is another advantage of the doubly stochastic method as the assignment, despite being coded in Python, runs in 12 hours 35 minutes. We estimate that using link elimination, which cannot be applied to multiple origin-destination pairs at a time, would take about 118 days.

### Avenues for Future Research

We tried to validate the trip assignment agains intersection counts, with poor results. We believe that the San Francisco model's weaknesses with respect to bicycles further up the chain result in poor trip tables being input to the assignment. Specifically, the mode choice utility specification for the bicycle alternative is currently a simple linear function of distance, ignoring the benefits of bicycle facilities and the dissuasive effects of hills.

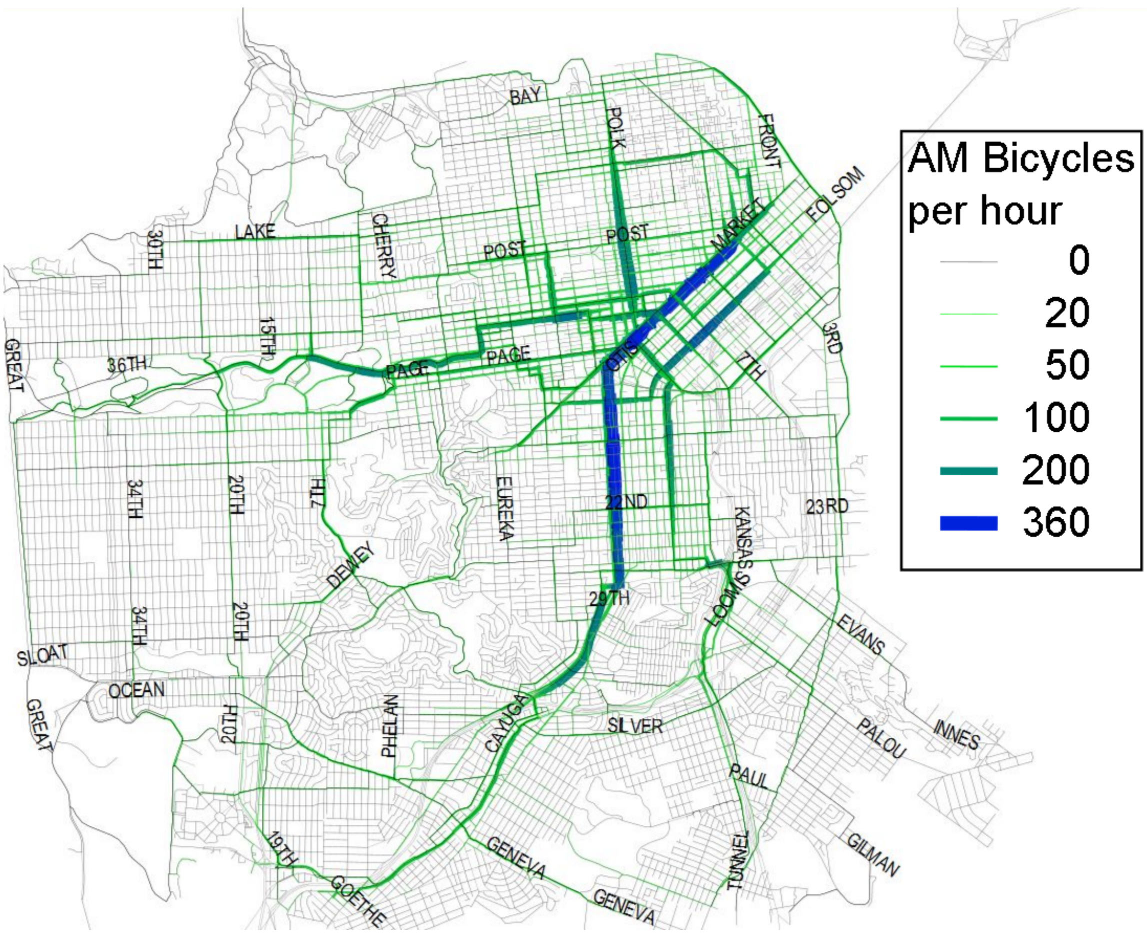



Figure 5: AM trip assignment

Therefore, we would like to feed the route choice logsums back to the mode choice model. To make this feedback possible, further research is needed into the functionality of, and necessary adjustments to, the logsum as a welfare evaluation measure under the presence of variably-sized choice sets with correlated random utilities. Once this methodological barrier is resolved, reflecting the attractiveness of route alternatives in the mode choice model would greatly enhance the responsiveness of the mode choice model to network conditions and, to our knowledge, provide the first conclusive evidence regarding the influence (or lack therof) of investments in bicycle infrastructure on mode choice. Given our result that infrequent cyclists value bike lanes more than frequent cyclists, we expect the influence to be significant.


## Acknowledgements

We thank Matt Paul, the programmer of CycleTracks; Michael Schwartz, the coordinator of the data collection effort; the San Francisco and Bay Area Bicycle Coalitions for promoting the study; Nadine Schüssler and Kay Axhausen for their GPS processing code; and Robert Cervero and Joan Walker for their comments. Mr. Hood is partially supported by a Graduate Research Fellowship from the National Science Foundation.


## References


Aultman-Hall, L., Hall, F. & Baetz, B. (1997), 'Analysis of bicycle commuter routes using geographic information systems: implications for bicycle planning', *Transportation Research Record: Journal of the Transportation Research Board* **1578**, 102–110.

Ben-Akiva, M. & Bierlaire, M. (1999), Discrete choice methods and their applications to short term travel decisions, *in* R. Hall, ed., 'Hanbook of Transportation Science', Kluwer Academic Publishers, Norwell, MA, chapter 2, pp. 5–34.

Bierlaire, M. (2003), BIOGEME: A free package for the estimation of discrete choice models, *in* 'Proceedings of the 3rd Swiss Transportation Research Conference', Ascona, Switzerland.

Bovy, P. & Fiorenzo-Catalano, S. (2007), 'Stochastic route choice set generation: behavioral and probabilistic foundations', *Transportmetrica* **3**, 173–189.

Bradley, M. (1988), 'Realism and adaptation in designing hypothetical travel choice concepts', *Journal of Transport Economics and Policy* **22**(1), 121–137.

Charlton, B., Schwartz, M., Paul, M., Sall, E. & Hood, J. (2010), CycleTracks: a bicycle route choice data collection application for GPS-enabled smartphones. Presented at the 3rd Conference on Innovations in Travel Modeling, a Transportation Research Board Conference, Tempe, AZ, May 10–12, 2010.

Forester, J. (1994), *Bicycle Transportation*, 2nd edn, The MIT Press, Cambridge.

Hagberg, A., Schult, D. & Swart, P. (2008), Exploring network structure, dynamics, and function using NetworkX, *in* G. Varoquaux, T. Vaught & J. Millman, eds, 'Proceedings of the 7th Python in Science Conference (SciPy2008)', Pasadena, CA, pp. 11–15.

Hunter, W., Stewart, J. R. & Stutts, J. (1999), 'Study of bicycle lanes versus wide curb lanes', *Transportation Research Record: Journal of the Transportation Research Board* **1674**, 70–77.

Krizek, K., Barnes, G., Poindexter, G., Mogush, P., Thompson, K., Levinson, D., Tilahun, N., Loutzenheiser, D., Kidston, D., Hunter, W., Tharpe, D., Gillenwater, Z. & Killingsworth, R. (2006), *Guidelines for Analysis of Investments in Bicycle Facilities*, Transportation Research Board, Washington, DC. NCHRP Report 552.

Menghini, G., Carrasco, N., Schüssler, N. & Axhausen, K. (2009), 'Route choice of cyclists: discrete choice modelling based on GPS-data', *Arbeitsberichte Verkehrs- und Raumplanung* **544**. IVT, ETH Zürich, Zürich.

Morpace International (2002), *Bay Area Travel Survey 2000: Final Report*, Metropolitan Transportation Commission, Oakland, CA.

Outwater, M. & Charlton, B. (2006), The San Francisco model in practice: Validation, testing, and application, *in* 'Innovations in Travel Demand Modeling: Summary of a Conference', Vol. 2, Transportation Research Board, Washington, DC.

Prato, C. & Bekhor, S. (2007), 'Modeling route choice behavior: How relevant is the composition of choice set?', *Transportation Research Record: Journal of the Transportation Research Board* **2003**, 64–73.

San Francisco Municipal Transportation Agency (2009), *San Francisco Bicycle Plan*, San Francisco, CA.

Schüssler, N. & Axhausen, K. (2009*a*), 'Map-matching of GPS traces on high-resolution navigation networks using the multiple hypothesis technique', *Working paper* **568**. IVT, ETH Zürich, Zürich.

Schüssler, N. & Axhausen, K. (2009*b*), 'Processing raw data from global positioning systems without additional information', *Transportation Research Record: Journal of the Transportation Research Board* **2105**, 28–36.

Schüssler, N., Balmer, M. & Axhausen, K. (2009), 'Route choice sets for very high-resolution data', *Arbeitsberichte Verkehrs- und Raumplanung* **567**. IVT, ETH Zürich, Zürich.

Sener, I., Eluru, N. & Bhat, C. (2009), 'An analysis of bicycle route choice preferences in Texas, US', *Transportation* **36**, 511–539.

Spellucci, P. (1988), 'An SQP method for general nonlinear programs using only equality constrained subproblems', *Mathematical Programming* **82**, 413–448.

Wilkinson, W., Clarke, A., Epperson, B. & Knoblauch, R. (1994), *The effects of bicycle accomodations on bicycle/motor vehicle safety and traffic operations*, National Technical Information Service, Great Falls.