NEXT DIRECTION ROUTE CHOICE MODEL FOR CYCLIST USING PANEL DATA
Melvin Wong, Polytechnique Montréal
Bilal Farooq, Polytechnique Montréal
Guillaume-Alexandre Bilodeau, Polytechnique Montréal

Introduction

Travel demand forecasting has been one of the major areas of interest in transportation research. In many urban cities in North America, bike sharing is a popular and heavily used mode of transport for many commuters downtown [1], [2]. When planning new infrastructures and facilities, these bike sharing services need to decide how to best improve level of service, user satisfaction and reliability. In literature on bike sharing models, the main objective is to describe a model which will explain the extent the relation between trip, environment variables and demand, such that the goal will be to improve transport operations and infrastructure design of new bike sharing services. Some papers have investigated the planning of new bike stations, impact on public transport and data analysis of GPS trackers in bikes [1], [3], [4].

User satisfaction is important in operating a bike sharing service, considering that demand can change over time and across individuals. An dynamic model can address these factors by having additional parameters which considers time and individual heterogeneity. We can leverage on data-driven dynamic models which recently have been shown to have strong performance in adapting to temporal and individual variations over static models [5], [6]. This raises new questions on how researchers should approach the new specifications problem. The key challenge here is finding the key relation between the heterogeneity and the short term travel dynamics and interactions. Panel data can provide such information on the observed choices. The implications of our study is highly relevant towards future transportation research, as we expect data-driven technology to be a key tool for the success of future transportation services [7], [8], [9].

Next-direction route choice model is based upon a microscopic level, near-future prediction method which aims to forecast the immediate next-step choice in a sequence of link-to-link trip data, rather than optimizing downstream route-to-destination utility. Our approach by itself does not look at the final destination, therefore, knowledge of origin-destination information is not required. This is the main advantage over traditional link-based route choice models. Another problem with link-based route choice models is that optimizing a route deep into a large tree network requires large computational effort because of the exponential growth of alternatives when we forecast a deeper network [10]. We can avoid this situation by considering only the adjacent links using a next-step choice method. A recent paper on predictive driving does a similar method using near-future prediction of next-direction probabilities looked into using sensory data from in-car and external cameras [11]. The paper describes a deep microscopic level approach which uses driver movements but is similar to our approach of forecasting near-future decisions. The focus of our work will look into the operational level component where we can determine level of demand across the network from predicted next-direction choices.

In this paper, we present our results from our dynamic mixed logit route choice model based purely on trip characteristics using the next-direction method, accounting for mixed-effects of individual heterogeneity observed from panel data and serial correlation of sequential observations. This paper will highlight the process in deciding choice set for estimation, explaining our dynamic mixed logit model and report results from our dataset. We will conclude with several hypothetical use for our forecasting
methods in travel models and also implications in transportation policy. Our research goal is to develop a framework for a predictive bike sharing load balancing application by observing future trajectories from our predictive model. We can forecast the travel behaviour of cyclists ahead of time and manage the transport of bikes from station to station in a more efficient and intuitive manner.

Literature review

Explanatory variables
The relationship between trip-level attributes including stated preference variables such as trip purpose and behaviour of travellers are commonly discussed in bicycle route choice model literature. A important decision impacting cycling behaviour is bicycle infrastructure and facilities in urban cities [12], [13]. The authors have found that these variables influence the travel habits of cyclists, indicating that cyclists have a stronger preference for dedicated bike facilities. At the individual level however, it becomes more difficult to analyse each person's taste and preference regarding trip decisions and interactions with the environment. Some studies have suggested that preferences of bike lanes, demographics and rider's skill [13], [14], [15]. Accessibility and commute time, urban spatial structure and socio-economic levels have also been discussed in travel behaviour modelling [8], [16], [17].

Panel Data
Time dependent variables, or dynamics, can be ideally captured by panel data [18], specifically when individuals are observed over a period of time [7]. Some authors have proposed using panel data as a means of detecting change in behaviour over time and using past choices to alter current and future choice probability [7], [9], [19]. This study of the dynamics in travel pattern allows one to control for unobserved variables, for instance, change in habits over time, that may possibly affect the choice outcome.

Dynamic Models
The decision behaviour of cyclists is based on the individual perception of route characteristics, taste heterogeneity and network rules. In addition, decisions made in the past will also affect current and future decisions. To account for this phenomena, we introduce dynamics into our route choice model. The most direct way of building a dynamic route choice model is to define a sequence of choices for every set of trips. Each choice reflects the explanatory variables at each interval. There are two forms dynamic modelling methods. First, a stochastic dynamic mixed model [9], where dynamics taste heterogeneity is estimated for every link in the trip and an alternative is selected for each link based on the previous choice. The other form of dynamic model is a recursive dynamic logit [20] where the goal is to maximize downstream utility towards a destination as a set of links (paths) recursively and a single path alternative is selected based on the highest utility factor of all possible paths.

Next-direction route choice model formulation

The next-direction model has two types of parameters: (i) static parameters $\beta$ for static explanatory variables and (ii) dynamic parameters $\gamma$ to describe past choice variables. When the mixed-effects are included into the transition parameters, we parametrise past choice variables with a normal distribution $\gamma \sim N(\beta, \sigma)$ with a mean $\beta$ and variance $\sigma$ to be estimated. The overall parameters to estimate for each next direction outgoing link $v \in R$ are: $\Theta_i = \{\beta, \beta, \sigma | v \in R\}$.

In a standard multinomial logit model, let $U_{nit}$ be the utility for a cyclist $n$ to take link direction alternative $i$ at time step $t$. Each time step has its own observed utility $V_{nit}$ and an unobserved component $\varepsilon_{nit}$:

$$U_{nit} = V_{nit} + \varepsilon_{nit}$$  (1)
where:

\[ V_{nitt} = \beta x_{niti} + \gamma_{niti} y_{niti(t-1)} \]  

(2)

The unobserved component \( \varepsilon_{niti} \) are assumed to be independent and iid extreme value distributed. \( \beta \) is a vector of parameters for time independent static explanatory variables \( x_{niti} \) and \( \gamma_{niti} \) is the vector of parameters for the time and individual dependent past choice variables \( y_{niti(t-1)} \) at the previous time step. Using the random parameters specification, the utility \( U_{niti} \) including the mixed-effect component of the dynamics, can be specified as:

\[ U_{niti} = \beta x_{niti} + \gamma_{niti} y_{niti(t-1)} + \varepsilon_{niti}; \quad \gamma_{niti} \sim \mathcal{N}(\hat{\beta}, \sigma) \]  

(3)

The probability of choosing outgoing link \( r \) for each link \( i \in \mathcal{C} \) where \( \mathcal{C} \) is the set of links in each trip at time step \( t \), \( P_i(r|\mathcal{C}, \beta, \gamma) \) with the availability of alternatives given as \( \delta \), can be expressed in a logit form:

\[ P_i(r|\mathcal{C}, \beta, \gamma) = \frac{\delta_i e^{\beta x_{niti} + \gamma_{niti} y_{niti(t-1)}}}{\sum_{j \in \mathcal{C}} \delta_j e^{\beta x_{niti} + \gamma_{niti} y_{niti(t-1)}}} \]  

(4)

The dynamic choice probability of the next link \( i \), \( P_{niti}(i|\mathcal{C}) \) for each individual \( n \) at time step \( t \) can be defined as:

\[ P_{niti}(i|\mathcal{C}) = \int_\beta \prod_{i \in \mathcal{C}} P_i(r|\mathcal{C}, \beta, \gamma) f(\beta, \sigma) d\beta \]  

(5)

Finally, the simulated maximum likelihood estimation function of the probability (5) is given as:

\[ SLL = \sum_{n=1}^{N} \sum_{t=1}^{T} \ln P_{niti}(i|\mathcal{C}) \]  

(6)

Since there is no closed form solution available, the log-likelihood probability function is estimated with Monte Carlo simulation with a Modified Latin Hypercube Sampling (MLHS) draws method using BIOGEME software to estimate parameters [21].

**Methodology**

To enable forecasting with panel data, there are a number of issues which need to be addressed. The major problem is the serial correlation of the panel data and taste heterogeneity between individuals. The problem is aggravated by the fact that individuals may change its decision choice over short-term changes in environment conditions and situations. Next, if we assume that final destination is unknown, real-time forecasting is difficult as there might be an infinite number of possible route alternatives. Having sufficient quantity and quality data is one reason that large data collecting tools, for example GPS is particularly suitable for trip data gathering because of the streaming nature of data. However, with GPS, we can only describe the characteristics of the trip (speed, duration, direction etc.) and not individual characteristics that are more commonly used in analysis.

To explain how we derived our choice set, we first classify direction of travel as a heading direction based on the compass principal directions: \{SW, W, NW, N, NE, E, SE, S\}. Following which, each link taken
in the path $i$ is given a set $\mathcal{R}$ of outgoing link alternatives $r$, given the link availability $\delta$, of outgoing links from the previous connected link. The objective is to choose one of the direction heading the traveller will visit. Next, dynamic component (action) is added for each transition from link to link, where the dynamic component is based on prior link (state) to next link (next state). Figure 1 is a simple illustration of a dynamic link-to-link network. The choices at each trip link can be defined as an output of static and dynamic variables $v_i$, $a_n$ respectively, shown in Figure 2. For every link, it takes inputs from the observed static and dynamic variables and generates a transition action and a choice alternative. With this framework, our next-direction choice model will be able to capture both the static explanatory variables as well as the past choice time and individual dependent variables with each trip segment.

![Figure 1: Example of a next-direction network with turn dynamics. The dynamics are based on the prior link choice (state) taken by the traveller.](image1)

![Figure 2: Example of sequence dynamics captured by static and dynamic variables. $s_1, \ldots, s_n$ represents observed choice decisions, $v_1, \ldots, v_n$ represents static explanatory variables and $a_1, \ldots, a_n$ represents dynamic variables based on past choices. Dynamic variables are specified as a mixed logit with variance parameter $\sigma$.](image2)

**Case study**

Trip data was collected from cyclists in Toronto, Canada using a GPS tracking smartphone application. There were a total of 61870 complete trips recorded between May 2014 to January 2015. For each state, a cyclist chooses one of seven different action choices {Straight, Left turn, Right turn, Near left turn, far left turn, near right turn, far right turn}. The dynamic variables are captured as an action sequence generated from the previous observation and chosen alternatives.

From our sample data we obtained, we see a strong dominance for $S$, $W$, and $E$ heading directions. This preference is likely due to the grid layout of arterial roads in the city of Toronto and Southbound trips towards the CBD. For building and testing of our model, we divided the trips into 70% for estimation and 30% for validation. Each trip level data contains information on time, link segment, speed and direction of travel. We tested our next-direction route choice model estimated with a sample of approximately 45000 trips with an average of 23 link instances per trip. The selected variables include speed, speed limit of link, volume of link and length of link, 4 road type dummy (Arterial, Collector, Local and Trail) and 3 time-of-day dummy (Morning: 8am – 12pm, Afternoon: 12pm – 4pm and Evening: 4pm – 8pm). The summary of variables are shown in Table 1. The travel heading $S$ was taken as a reference; $ASC_S = 0$. We also made several assumptions: (i) Origin and Destination location and departure time is independent of choice, and (ii) We set initial direction choice to be zero and dynamics are based only on previously connected links.
Model results

Table 1 presents the estimated parameter results and comparison of next-direction model performance for a dynamic mixed effect model with panel effects, a static model without dynamics and a dynamic exogenous model where dynamics are not modelled as mixed effect variables. The log-likelihood ratio test is used to compare models, and the estimated dynamic mixed model is shown to have a statistical improvement over the dynamic exogenous model and the static model. Cyclists have a preference for paths frequently used by other cyclists as indicated by the reported parameters showing negative sign for short links and positive sign for speed limit and volume of traffic. From our results, all dynamic variables are particularly significant with higher preference for left and right turns over going straight, however left and right turn $\sigma$ are not significant, meaning that these parameters are not influenced by stochastic variability across individuals or time.

Estimation results also indicate that the dynamic mixed model improves fit over the static and dynamic exogenous models. Also, to note that all 3 models have similar coefficients. When comparing the dynamic mixed model with the dynamic exogenous model, we see that some of the $\sigma$ parameters has a high significance, validating that some mixed effects of dynamics have high influence on the estimated model. The $\beta$ and $\sigma$ coefficients in Table 1 are statistically significant at 95% indicated by the t-test value.

A positive result is indicated on the confusion matrix diagonal where the simulated choice matches the actual choice (Figure 5). We use accuracy to measure the performance of our model. Number of simulated observations are given in the last column. As shown in figure 5, the dynamic mixed model performs better than the other two models, on average, the normalized accuracy is 65.30% compared to 63.83% for the static model and 64.89% for the dynamic exogenous model. Normalization is performed to account for the systematic bias on each alternative. We note that for alternative 4, corresponding to direction N,
has low accuracy compared to the other alternatives, the reason for this might be caused by a biased dataset. However, the improvement in overall accuracy when we account for mixed-effect indicates that dynamics are influenced by external factors not observed by the observed utility.

**Discussion on the benefits of predictive models**

In the present analysis, we attempt to use our estimated models to forecast route choices based on the next-direction alternative. A test scenario is set up where a sample validation set containing approximately 15000 trips is used to estimate the accuracy of the models. First the static and dynamic exogenous models are compared to the dynamic mixed model on single link accuracy via a simulation on the validation set.

Predictive models can be used to anticipate manoeuvres in real time to give transport operators a picture of near-future demand. This will help to plan and improve new infrastructure and facilities or provide smart demand-based services for travellers such as forecasting bike station capacity in a short time span ahead. With richer data and more data integration from traffic sensors such as bike counters, we can further improve their performance. Our approach can be used for certain practical applications like service load balancing, demand forecasting and dynamic street pricing. Because our model is a short-term predictive model, we can adapt to changes in traffic situations more accurately than traditional demand choice models.

Another situation which predictive model is applicable is assisted driving. Autonomous driving has been a hot topic recently in many transportation and machine learning research fields. One possible step is to use our dynamic mixed model and apply a learning layer for example, a neural network, for adaptive parameter changes according to real time traffic interactions. This would allow autonomous vehicles to anticipate manoeuvres and alert the driver of possible dangers ahead or traffic congestion. In our future work, we propose to use a hybrid model which may give better predictive performance. A hybrid model will offer a way to change parameter values dynamically.

The results of this study show indications that panel data effect play an important role in modelling dynamics of discrete behavioural process. However, estimating parameters based on static discrete time points would not be sufficient to capture all unobserved components. Furthermore, a mixed-effect dynamic component improves estimation accuracy, however only slightly compared to a non-mixed dynamic model. This is because there are other serially dependent factors which can alter the behavioural dynamics that are not considered in this paper, for example: condition or type of road, by district, or proximity to landmarks. It is important to note that there are also other unobserved trip attributes which are not obtainable in our case study. Our model requires further improvement as we see that accuracy prediction is not consistent across all alternatives. One possible reason might be due to the alternative bias, where N and NE bound directions has the least number of observations corresponding to the lowest accuracy. Our model may be biased towards directions which are more frequent. Also, due to the grid-like network of our case study area, certain directions for example, NW, NE, SE and SW are less likely to get chosen over others.

Indeed it seems that panel data establishes a suitable practical approach to modelling a dynamic forecasting model. Additional factors that need to be considered for improvement include spatial segmentation by road configuration or by districts and multi-modal interaction with other road vehicles and pedestrians. In developing a strategy for cyclist travel forecasting, dynamics of link-to-link choices should be considered in addition to attributed related to the network. Other factors, for instance availability of dedicated bike paths, location of bike parking or bike sharing stations can be further explored as a mixed-effect dynamic variables which are not discussed in depth in this paper.
Figure 4: Results of our predictions showing accuracy for each heading direction. Simulated predictions are on rows and actual values are on columns. Last column is the number of simulated observations. Right: Static model; Bottom left: Dynamic exogenous model; Bottom right: Dynamic mixed model.

Table 1: Estimation results. Significant parameters are shown.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Static Model</th>
<th>Dynamic Exogenous</th>
<th>Dynamic Mixed</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASC_SW</td>
<td>-0.131</td>
<td>0.0263</td>
<td>-0.283</td>
</tr>
<tr>
<td>ASC_E</td>
<td>-0.113</td>
<td>0.0303</td>
<td>-0.112</td>
</tr>
<tr>
<td>ASC_S</td>
<td>-0.310</td>
<td>0.0198</td>
<td>-0.432</td>
</tr>
<tr>
<td>ASC_NW</td>
<td>-2.97</td>
<td>0.0300</td>
<td>-9.896</td>
</tr>
<tr>
<td>ASC_SE</td>
<td>-0.507</td>
<td>0.0133</td>
<td>-38.19</td>
</tr>
<tr>
<td>ASC_NE</td>
<td>0.149</td>
<td>0.00416</td>
<td>35.63</td>
</tr>
<tr>
<td>ASC_SE</td>
<td>0.271</td>
<td>0.00806</td>
<td>33.60</td>
</tr>
<tr>
<td>ASC_S</td>
<td>0</td>
<td>0.00312</td>
<td>-8.19</td>
</tr>
<tr>
<td>β_SPEED</td>
<td>0.00127</td>
<td>1.74e-05</td>
<td>-73.58</td>
</tr>
<tr>
<td>β_LENGTH</td>
<td>0.0815</td>
<td>0.000625</td>
<td>130.50</td>
</tr>
<tr>
<td>β_SPEEDLIMIT</td>
<td>0.00103</td>
<td>6.27e-06</td>
<td>165.00</td>
</tr>
<tr>
<td>β_VOLUME</td>
<td>0.929</td>
<td>0.0310</td>
<td>29.97</td>
</tr>
<tr>
<td>β_ARTERIAL_DUMMY</td>
<td>0.0297</td>
<td>0.00947</td>
<td>3.14</td>
</tr>
<tr>
<td>β_COLLECTOR_DUMMY</td>
<td>0.122</td>
<td>0.00941</td>
<td>13.00</td>
</tr>
<tr>
<td>β_LOCAL_DUMMY</td>
<td>-1.42</td>
<td>0.0310</td>
<td>-45.73</td>
</tr>
<tr>
<td>β_TRAIL_DUMMY</td>
<td>-0.188</td>
<td>0.0323</td>
<td>-5.82</td>
</tr>
<tr>
<td>β_MORNING_DUMMY</td>
<td>-0.313</td>
<td>0.0357</td>
<td>-8.77</td>
</tr>
<tr>
<td>β_EVENING_DUMMY</td>
<td>-0.445</td>
<td>0.0356</td>
<td>-12.49</td>
</tr>
<tr>
<td>β_LEFTTURN</td>
<td>0.861</td>
<td>0.0400</td>
<td>21.50</td>
</tr>
<tr>
<td>β_FAR_LEFTTURN</td>
<td>0.241</td>
<td>0.0350</td>
<td>6.78</td>
</tr>
<tr>
<td>β_NEAR_LEFTTURN</td>
<td>1.11</td>
<td>0.0491</td>
<td>22.65</td>
</tr>
<tr>
<td>β_RIGHTTURN</td>
<td>0.601</td>
<td>0.0396</td>
<td>15.18</td>
</tr>
<tr>
<td>β_NEAR_RIGHTTURN</td>
<td>1.51</td>
<td>0.0354</td>
<td>42.61</td>
</tr>
<tr>
<td>β_FAR_RIGHTTURN</td>
<td>0.943</td>
<td>0.0513</td>
<td>18.39</td>
</tr>
<tr>
<td>β_STRAIGHT</td>
<td>0.462</td>
<td>0.0317</td>
<td>14.59</td>
</tr>
<tr>
<td>σ_FAR_LEFTTURN</td>
<td>-1.35</td>
<td>0.0461</td>
<td>-29.33</td>
</tr>
<tr>
<td>σ_NEAR_LEFTTURN</td>
<td>2.56</td>
<td>0.110</td>
<td>23.19</td>
</tr>
<tr>
<td>σ_NEAR_RIGHTTURN</td>
<td>1.75</td>
<td>0.0376</td>
<td>46.52</td>
</tr>
<tr>
<td>σ_FAR_RIGHTTURN</td>
<td>-3.44</td>
<td>0.153</td>
<td>-22.52</td>
</tr>
<tr>
<td>σ_STRAIGHT</td>
<td>0.0513</td>
<td>0.0513</td>
<td>5.59</td>
</tr>
</tbody>
</table>

Null Log likelihood: -748667.232
Final Log likelihood: -543832.689
References