Analysis of Key Route Attributes for Route Choice Model Estimation Based on GPS Data

Anna Danielsson*[1], David Gundlegård[1], and Clas Rydergren[1]

[1]Department of science and technology, Linköping University, Sweden

SHORT SUMMARY

Efficient traffic control requires an understanding of vehicle flows in the road network, where two important components are demand and route choice. The aim of this study is to use GPS data to identify key attributes explaining the route choice observed in data. The data set consisted of about 150,000 trips and was divided into a training dataset and a test dataset. The two datasets were compared and experiments show that the routes used are similar. Discrete route choice models were estimated with a route set of observed routes in the training data. The best model showed a result of 73 % correct predictions and a route choice distribution error of 19 %. Furthermore, the result indicated that simplicity of the route is more important than the travel time.

Keywords: Discrete choice, GPS data, Route attributes, Route choice

1. INTRODUCTION

Transportation accounts for approximately a fourth of the global carbon dioxide emissions (IEA, 2019). One contributing factor is congestion, which occurs when the number of vehicles on a road reaches the capacity limit (Treiber & Kesting, 2013). While a straightforward solution would be to add more capacity to the road network, this would not be an efficient use of resources since it induces new traffic demand (Melo, Graham, & Canavan, 2012). Thus, controlling and managing the traffic given the current network would be a cost efficient way of reducing congestion. The first step for efficient traffic control is to understand the flows, where two important components are knowledge about demand and route choice. This study aims to use GPS data to identify key attributes that can explain the route choice.

There have been various studies examining route choice with different purposes in the past, many using discrete choice models. Hess et al. (2015) and Montini, Antoniou, and Axhausen (2017) both develop multinomial logit models based on GPS trajectories, with focus on heavy goods vehicles and public transport. Dhakar and Srinivasan (2014) develop a path size logit model and focus on finding key attributes by combining GPS traces with a travel survey. Furthermore, Duncan et al. (2021) integrated a path size logit model with a bounded choice model to overcome the problems of route correlations and unrealistic routes among the alternatives. They emphasize the importance of having a route choice set that includes only the routes that are being used to get a realistic result. Qin et al. (2019) propose a taxi trip choice model for anomaly detection, and they handled the choice set size issue by utilizing the historically observed routes as the choice set. Fosgerau, Frejinger, and Karlstrom (2013) and Mai, Yu, Gao, and Frejinger (2021) both suggest a recursive logit model that formulates the route choice model as a sequence of link choices where the decision maker chooses the utility-maximizing outgoing link at each node. With this approach, there is an infinite number of alternatives in the choice set. Yao and Bekhor (2020) deal with the choice set issue by sampling a choice set from route characteristics clusters. Axhausen and Schüssler (2009) discuss the similarity between the alternatives in addition to the choice set size issue. They evaluate both different choice set generation algorithms and different formulations of the similarity factor. Their conclusion is that a performance optimised Breadth First Search on Link Elimination together with a choice set reduction procedure and a road type specific path size factor was the best in their study.

Regardless of the application of the route choice model and the choice set size, travel time is a common attribute to describe the utility of a route. It could be the free flow travel time, measured travel time, or travel time calculated as a function of demand and capacity. Table 2 shows additional attributes that were found in road transport literature using passive data.

Table 1: Attributes in a route choice model

Attributes	Occurrences
Travel time	Hess et al., 2015; Dabbas et al., 2021; Montini et al., 2017; Deng et al., 2021; Dhakar & Srinivasan, 2014; Duncan et al., 2021; Fosgerau et al., 2013; Axhausen & Schüssler, 2009; Mai et al., 2021; Yao & Bekhor, 2020; Bovy et al., 2008
Path size/commonality factor/ correlation factor/link size	Hess et al., 2015; Montini et al., 2017; Duncan et al., 2021; Fosgerau et al., 2013; Axhausen & Schüssler, 2009; Mai et al., 2021; Bovy et al., 2008
Route length	Deng et al., 2021; Dhakar & Srinivasan, 2014; Duncan et al., 2021; Yao & Bekhor, 2020; Bovy et al., 2008
Road type	Montini et al., 2017; Deng et al., 2021; Axhausen & Schüssler, 2009; Yao & Bekhor, 2020; Bovy et al., 2008
Number of turns and signals	Montini et al., 2017; Deng et al., 2021; Dhakar & Srinivasan, 2014; Fosgerau et al., 2013; Yao & Bekhor, 2020
Number of crossings/intersections/ links	Deng et al., 2021; Dhakar & Srinivasan, 2014; Fosgerau et al., 2013; Mai et al., 2021; Yao & Bekhor, 2020
Speed (average or maximum)	Dhakar & Srinivasan, 2014; Yao & Bekhor, 2020; Bovy et al., 2008
Route delay	Yao & Bekhor, 2020; Bovy et al., 2008
Route cost	Hess et al., 2015; Yao & Bekhor, 2020
Euclidean distance in OD-pair	Yao & Bekhor, 2020; Dhakar & Srinivasan, 2014
Path size depending on road type	Axhausen & Schüssler, 2009
Number of consecutive motorway links (plausibility)	Axhausen & Schüssler, 2009
Percentage of route in city center	Yao & Bekhor, 2020
Variance of travel time	Mai et al., 2021

Based on observed trajectory data from Stockholm, we create a route set and estimate discrete route choice models to fulfil the aim.

2. METHODOLOGY

A dataset of GPS trips were divided into a training dataset used for model estimation and a test dataset used for validation. The route set to be used in the discrete choice model were the historically observed routes in the training dataset, (e.g. Qin et al., 2019). Based on the training dataset, logit models were estimated to predict route choice. The probability of alternative i being chosen out of the route set C is calculated using Equation 1.

$$P(i|C) = \frac{e^{V_i}}{\sum_{j \in C} e^{V_j}} \tag{1}$$

where V_i is the utility of alternative i based on a number of attributes. The general structure of the utility function is $V_i = \sum_k \beta_k x_{ik}$, with coefficients β_k and values x_{ik} for each attribute k and alternative i. Some attributes are related to the route characteristics and does not vary between the travelers, such as route length and road type. Others are dependent on the driver and what time of the day the trip occurred, such as travel time and delay. Since all routes were not observed during all times, travel time was calculated using the link mean travel time over four time intervals. A link travel time variance was also calculated and summarised to a route variance.

Models were estimated using maximum likelihood and performance statistics for evaluation were log-likelihood, BIC, hitrate and RMSE of predicted and observed route probabilities. The log-likelihood increases with more attributes, possibly resulting in overfitting the model. The BIC value attempts to avoid that by introducing a penalty for the number of attributes. When observed route choice had the largest predicted route probability, it was considered a hit, and dividing the number of hits with the number of observations gives a hitrate (Bovy et al., 2008). One model was estimated based on forward stepwise selection (FSS), where one attribute at a time is added to the model based on some criteria, e.g. BIC (James et al., 2017). The hitrate and RMSE of route probabilities were computed also for the test data to get the test error as validation.

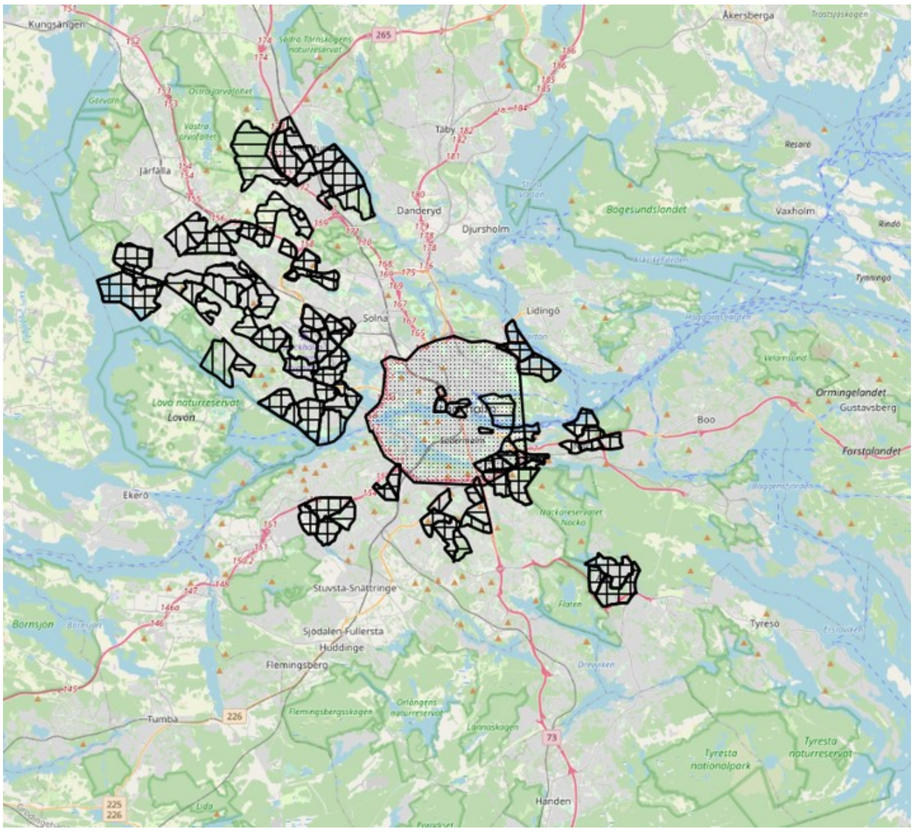

Figure 1: OD-pairs and city area of Stockholm.

3. RESULTS AND DISCUSSION

A dataset of INRIX GPS trajectories from Stockholm was used to estimate the route choice models. The dataset was collected during 5 weeks and consist of 593,153 trips, with a penetration rate of 1-2%. More information about the dataset is presented in Ahlberg et al. (2021). A *trip* is a series of links used by a traveler building up a *route* between origin and destination. The origins and destinations were aggregated over 558 zones, and the 124 origin-destination pairs (OD-pairs) with the largest number of observations were used in the analysis (Figure 1). To increase the number of observations, all trips that pass by the origin and destination zones are included, not only the ones starting and ending there.

The first two weeks of the dataset (weekends excluded) constitutes the training dataset (65,784 trips), while the following two weeks constitutes the test data (83,800 trips). Figure 2 shows example OD-pairs with training data routes in blue and test data routes in orange. The most used routes in each OD-pair primarily uses the major roads. Routes more than 50% longer than the shortest route in each OD-pair were filtered out. The left part of Figure 3 illustrates the number of routes in the two datasets (in total 1303 in training data and 1557 in test data). The right part of Figure 3 presents the percentage of routes in the training data that are used also in the test data when routes with less observations than a threshold are excluded. When all observations are included (threshold = 0), around 60 % of the routes in the training data are used in the test data while more than 90 % of the routes are used when all routes only used once are excluded (threshold = 1). Although excluding routes gives a more stable route set, there is a risk of excluding routes rarely used because of the low penetration rate. When the exclusion threshold is one, two and three, the number of routes decreases with 46 %, 45 % and 40 % respectively compared to all routes in the training data.

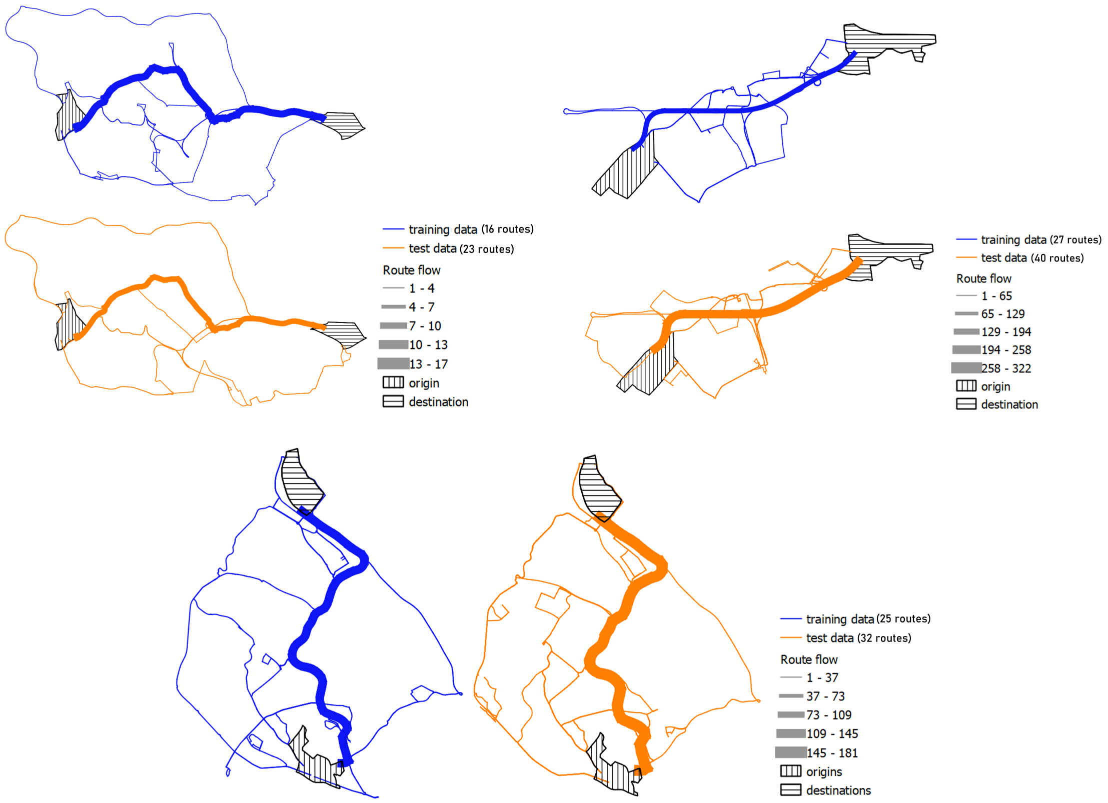

Figure 2: Routes training data (blue), routes test data (orange) of three example OD-pairs. The width shows the number of observations of each route.

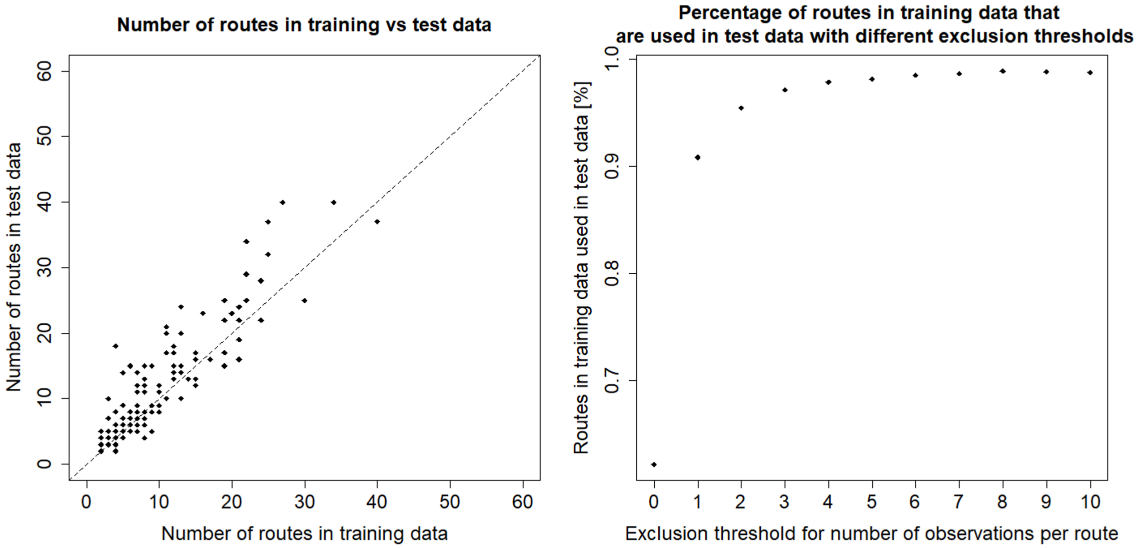

Figure 3: Number of routes in the training dataset vs number of routes in the test data for the 124 OD-pairs with highest demand to the left, and similarity between the datasets to the right.

Table 2: Route attribute statistics, difference within OD pairs

Attribute	Average	min diff	max diff	average diff	s.d. diff
ttmean [min]	1.18	0.24	29.94	4.51	4.86
ttfree [min]	0.81	0	13.47	1.87	2.38
delay [share]	0.47	0.17	62.30	3.45	7.07
variance [min]	21.70	1.62	1,277.20	273.89	290.71
rlength [km]	0.78	0	16.42	1.30	1.97
numlinks	5.48	0	138	15.78	19.47
p_city [share]	0.33	0	1	0.07	0.20
p_major_roads [share]	0.75	0	1	0.47	0.43
cf	0.74	0	2.62	0.85	0.64

Based on the overview of attributes in the Introduction and the available data, the attributes included in this study were mean travel time in minutes (ttmean), free flow travel time in minutes (ttfree), route length in kilometers (rlength), relative delay (ttmean - ttfree)/ttfree), route variance, number of links (numlinks), percentage length of trip within the city center defined in Figure 1 (p_city), percentage length of major roads in route (p_major_roads) and a commonality factor as Huang et al. (2018) (cf). The commonality factor is included to account for the overlap between the alternatives.

Route attribute statistics for the training dataset are shown in Table 2, presented as difference within the OD-pairs. The *Average* column presents the average over all routes in all OD-pairs. That the average ttmean is shorter than the average difference indicates that there are a lot of short routes, which could be an effect of that all routes passing by the OD-pairs are included (more trips passes by short OD-pairs than long ones). That is indicated also by the low averages of ttfree, rlength and numlinks.

Three route choice models and a null model for comparison were estimated. The nullmodel (mNull) does not include any attributes and divides the travelers equally over the alternative routes. The second model includes the most commonly used attributes in Table 2 (mTable). The third model was selected inspired by the the key attributes suggested in Yao and Bekhor (2020) (mYao). For the fourth model, the FSS method based on BIC was used to choose the attributes (mFSS). Figure 4, presenting the BICs of the best models for every number of attributes, suggests that up to eight attributes, the model gets better the more attributes included. The left part of the figure presents what attribute(s) to include with the fixed number of attributes. Thus, if only one attribute is included, the best one to include is p_major_roads. The lowest BIC value is obtained when 8 attributes are included. The somewhat surprising fact that ttmean was not included earlier in the model, could be a result of the route set generation process. All routes in the route set have been used at least once, thus at least one traveler considered it to be the best route. When routes are generated, as in conventional route set generation, unreasonable routes are included in the route set.

The resulting models are shown in Table 3, presenting the coefficient values with p-values indicated by asterisks. As expected, all models are better than the null model, and a low BIC value, a high hitrate and high log-likelihood indicate that mFSS is the best model for this dataset. Although, the RMSE is lower for mYao. For the mFSS model all coefficients are significant, and most of them have reasonable signs. ttmean, ttfree and delay all decrease the utility, which is reasonable. However, rlength increases the utility of the route, which could indicate that travelers take detours to use major roads. Thus, simplicity of the route seems to be more important than the distance, which explains why the coefficient for p_major_roads is positive and the coefficient for numlinks is negative. That a larger percentage of city center links increases the utility could

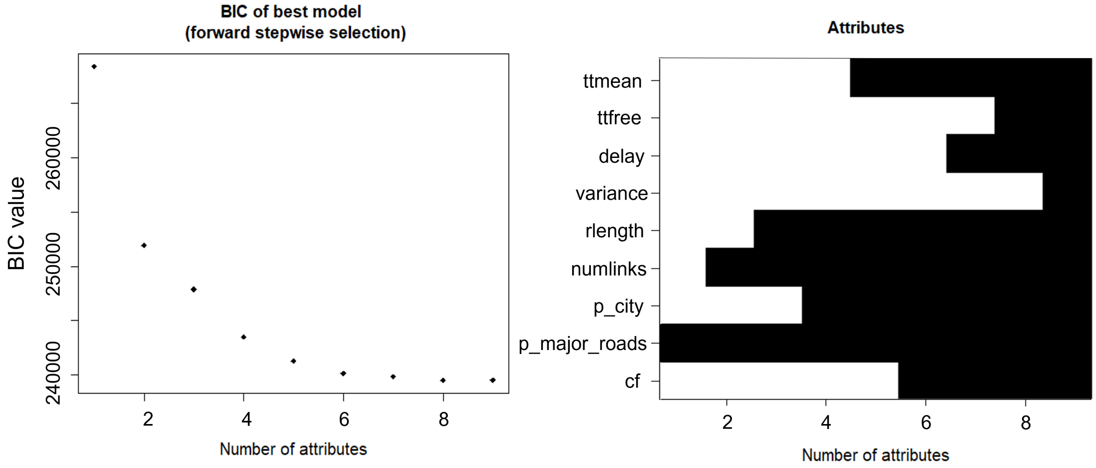

Figure 4: BICs of the best models for fixed number of attributes to the right. The result suggests that the more attributes there are the better. The attributes included in each model is shown in the left diagram.

be explained by the fact that one of the largest roads in Stockholm was included in the city center area. The coefficient for the commonality factor is negative, which is in line with the purpose of the attribute.

The hitrates and RMSE of the test data are only slightly worse, or even better than the training data metrics, which suggest a low risk of an overfitted model with good performance also on new data. The test hitrates for datasets with exclusion threshold one, two and three respectively are 83.19 %, 83.41 % and 85.08 %. Thus, the hitrate improves when more routes are excluded (e.g. Duncan et al., 2021).

Table 3: Model estimations with coefficients for each attribute and performance metrics for both training and test data.

Attribute	mNull	mTable	mYao	mFSS
ttmean		-0.40^{***}	-0.28^{***}	-0.13^{***}
ttfree				-0.73^{***}
rlength		0.95^{***}	0.89^{***}	1.30^{***}
p_major_roads		2.05^{***}	1.88^{***}	1.70^{***}
cf		-0.47^{***}		-0.38^{***}
p_city			1.69^{***}	1.51^{***}
delay			-0.16^{***}	-0.26^{***}
variance				
numlinks		-0.10^{***}	-0.14^{***}	-0.10^{***}
Training data				
BIC	309,899	244,833	240,932	239,512
Log Likelihood	−154,944	−122,390	−120,434	−119,713
Hitrate [%]	NA	68.77	69.02	75.50
Route probabilites [RMSE]	0.195	0.192	0.188	0.191
Observations	65,784	65,784	65,784	65,784
Test data				
Hitrate [%]	NA	68.08	69.19	72.58
Route probabilites [RMSE]	0.179	0.173	0.169	0.172
Observations	83,800	83,800	83,800	83,800
Note:	*p<0.1; **p<0.05; ***p<0.01 Table created with the Stargazer package in R Hlavac (2018).			

4. CONCLUSIONS

Observed GPS data was used to identify key route attributes for discrete route choice models. The literature review summarized common attributes to include and emphasized the importance of a reasonable route set. The result suggested that simplicity seems to be more important than travel time and distance, and the best model showed a test hitrate of 73 % and RMSE of 19 %.

The test hitrate increased from 73 % to 83 % when a more filtered route set was used (exclusion threshold of 1), which is in line with existing literature. Thus, the result suggests that building up the choice set of the historically observed routes is promising. However, the number of routes decreased with 46 %, risking exclusion of rarely used routes. Thus, there is a trade-off between model performance and route set size.

The results of the study together with a large increase in available GPS data indicates a promising future for improved data-driven route choice models. Improved route choice models are an important component of traffic management tools that enables more efficient actions for congestion reduction. As future work, it would be interesting to include more attributes and analyse how more data as well as different type of estimation methods affect the result.

ACKNOWLEDGMENTS

This work was supported by the Swedish Transport Administration [grant number TRV 2020/118663 and TRV 2018/132473].

REFERENCES

Ahlberg, J., Danielsson, A., Drageryd, L., Gundlegård, D., Ramsey, J., Sjöholm, A., & Sjöstrand, S. (2021). *Probedata: förstudie kring användning av gps-baserad probedata för skattning av hastigheter, länkflöden och ruttval* (No. TRV 2019/98384).

Axhausen, K. W., & Schüssler, N. (2009). Accounting for route overlap in urban and suburban route choice decisions derived from GPS observations. , 32 p. (Artwork Size: 32 p. Medium: application/pdf Publisher: ETH Zurich) doi: 10.3929/ETHZ-A-005916981

Bovy, P. H. L., Bekhor, S., & Prato, C. G. (2008, January). The Factor of Revisited Path Size: Alternative Derivation. *Transportation Research Record: Journal of the Transportation Research Board*, *2076*(1), 132–140. doi: 10.3141/2076-15

Dabbas, H., Fourati, W., & Friedrich, B. (2021). Using Floating Car Data in Route Choice Modelling - Field Study. *Transportation Research Procedia*, *52*, 700–707. doi: 10.1016/j.trpro.2021.01.084

Deng, Y., Yan, S., Hu, X., & Zhang, P. (2021, March). Mining Route Set Distribution Range and Affecting Factor Threshold Based on Global Positioning System Data. *Transportation Research Record: Journal of the Transportation Research Board*, 036119812199905. doi: 10.1177/0361198121999059

Dhakar, N. S., & Srinivasan, S. (2014, January). Route Choice Modeling Using GPS-Based Travel Surveys. *Transportation Research Record: Journal of the Transportation Research Board*, *2413*(1), 65–73. doi: 10.3141/2413-07

Duncan, L. C., Watling, D. P., Connors, R. D., Rasmussen, T. K., & Nielsen, O. A. (2021). A bounded path size route choice model excluding unrealistic routes: formulation and estimation from a large-scale GPS study. *Transportmetrica A: Transport Science*, 1–59. doi: 10.1080/23249935.2021.1872730

Fosgerau, M., Frejinger, E., & Karlstrom, A. (2013, October). A link based network route choice model with unrestricted choice set. *Transportation Research Part B: Methodological*, *56*, 70–80. doi: 10.1016/j.trb.2013.07.012

Hess, S., Quddus, M., Rieser-Schüssler, N., & Daly, A. (2015). Developing advanced route choice models for heavy goods vehicles using GPS data. *Transportation Research Part E: Logistics and Transportation Review*, *77*, 29–44. doi: 10.1016/j.tre.2015.01.010

Hlavac, M. (2018, May). *Package stargazer.* Comprehensive R Archive Network (CRAN). Retrieved from https://cran.r-project.org/web/packages/stargazer/index.html

Huang, Z., Huang, Z., Zheng, P., & Xu, W. (2018, May). Calibration of C-Logit-Based SUE Route Choice Model Using Mobile Phone Data. *Information*, *9*(5),

115. (Number: 5 Publisher: Multidisciplinary Digital Publishing Institute) doi: 10.3390/info9050115

IEA. (2019). *Transport: Improving the sustainability of passenger and freight transport.* Retrieved from https://www.iea.org/topics/transport

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2017). *An introduction to statistical learning with applications in r.* Springer.

Mai, T., Yu, X., Gao, S., & Frejinger, E. (2021, September). Routing policy choice prediction in a stochastic network: Recursive model and solution algorithm. *Transportation Research Part B: Methodological*, *151*, 42–58. doi: 10.1016/j.trb.2021.06.016

Melo, P. C., Graham, D. J., & Canavan, S. (2012, January). Effects of Road Investments on Economic Output and Induced Travel Demand: Evidence for Urbanized Areas in the United States. *Transportation Research Record*, *2297*(1), 163–171. (Publisher: SAGE Publications Inc) doi: 10.3141/2297-20

Montini, L., Antoniou, C., & Axhausen, K. W. (2017). Route and mode choice models using GPS data. , 14 p. (Artwork Size: 14 p. Medium: application/pdf Publisher: ETH Zurich) doi: 10.3929/ETHZ-B-000119554

Qin, G., Huang, Z., Xiang, Y., & Sun, J. (2019, January). ProbDetect: A choice probability-based taxi trip anomaly detection model considering traffic variability. *Transportation Research Part C: Emerging Technologies*, *98*, 221–238. doi: 10.1016/j.trc.2018.11.016

Treiber, M., & Kesting, A. (2013). *Traffic Flow Dynamics*. Berlin, Heidelberg: Springer Berlin Heidelberg. doi: 10.1007/978-3-642-32460-4

Yao, R., & Bekhor, S. (2020, December). Data-driven choice set generation and estimation of route choice models. *Transportation Research Part C: Emerging Technologies*, *121*, 102832. doi: 10.1016/j.trc.2020.102832