Performance Evaluation of Choice Set Generation Algorithms for Analyzing Truck Route Choice: Insights from Spatial Aggregation for the Breadth First Search Link Elimination (BFS-LE) Algorithm

Divyakant Tahlyan
Graduate Student
Department of Civil and Environmental Engineering
Northwestern University
2145 Sheridan Road, Evanston, IL 60208
Email: dtahlyan@u.northwestern.edu

Abdul Rawoof Pinjari*
Associate Professor
Department of Civil Engineering
Centre for infrastructure, Sustainable Transportation, and Urban Planning (C*i*STUP)
Indian Institute of Science (IISc)
Bangalore 560012, India
Ph: +91-80-2293-2043 Fax: +91-80-2360-0404
Email: abdul@iisc.ac.in

Word Count (excluding appendix): 13,346

Word Count (only appendix): 709

*Corresponding Author

Abstract

This paper presents a new approach to evaluate route choice set generation algorithms, which is based on comparing the algorithm-generated choice sets for an origin-destination (OD) location pair against the *portfolio* of observed routes for that OD pair. The approach offers the ability to evaluate both the generation of relevant routes that are typically considered by the travelers and the generation of irrelevant (or extraneous) routes that are seldom chosen. Although the approach is general enough for evaluation of most route choice set generation algorithms, this paper focuses on evaluating the performance of a breath first search link elimination (BFS-LE) approach to generating route choice sets for truck travel. To demonstrate the usefulness of the approach, route choice sets generated from the BFS-LE algorithm are evaluated against observed truck routes derived from large streams of truck-GPS traces in the Tampa Bay region of Florida. To do so, a systematic procedure is used to arrive at an appropriate combination of: (a) the spatial aggregation for the origin and destination locations and (b) the minimum number of trips to be observed between each OD location for evaluating algorithm-generated choice sets. Based on the evaluation, the paper offers guidance on effectively using the BFS-LE approach to maximize the generation of relevant routes. It is found that carefully chosen spatial aggregation can reduce the need to generate large number of routes for each trip. Further, estimation of route choice models and their subsequent application on validation datasets revealed that the benefits of spatial aggregation might be harnessed better if irrelevant routes are eliminated from the choice sets. Lastly, a comparative analysis revealed systematic differences in route characteristics of the relevant and irrelevant routes, an understanding of which can help eliminate the latter.

Keywords: *route choice, choice set generation, BFS-LE, truck-GPS data, spatial aggregation*

1. Introduction

Route choice set generation is an important precursor to analyzing travelers' route choice[1]. The route choice set for a given origin-destination (OD) location pair is a subset of feasible alternative routes offered by the transportation network between that OD pair. However, the number of feasible routes in real life networks is typically very large, computationally difficult to enumerate, not readily distinguishable from each other (due to overlaps), unknown to travelers, and varies substantially from one OD pair to another (Bovy, 2009). Therefore, extraction of the set of routes known to and potentially considered by travelers (which comprises the consideration set) (Hoogendoorn-Lanser, 2005) is a challenging task. A variety of different choice set generation algorithms have been used in the literature to generate route choice sets (Ben-Akiva et al., 1984; Bovy and Fiorenzo-Catalano, 2007; de la Barra et al., 1993; Prato and Bekhor, 2006; Rieser-Schüssler et al., 2013; Rieser-Schüssler and Axhausen, 2009). Most of these algorithms focus on generating alternative routes that are behaviorally realistic (for example, acyclic routes) and diverse (i.e., routes that do not overlap significantly), with a goal to maximize the generation of relevant routes that are likely to be chosen by travelers while reducing the generation of irrelevant routes that are not typically considered by travelers (for example, routes with large detours from shortest paths). As the composition of choice sets potentially can have a significant impact on model estimation and prediction (Bliemer and Bovy, 2008; Prato and Bekhor, 2007), evaluation of the generated choice sets is an important step prior to using them for route choice analysis.

A widely used approach to evaluate route choice set generation algorithms is to measure the extent to which the generated choice sets include the observed travel routes. This approach operates at a *trip level*, where for each observed trip, it is assessed whether the generated route choice set includes the observed route within a certain tolerance level (Bekhor et al., 2006; Prato and Bekhor, 2007). The proportion of observed trips for which the generated choice sets include

[1] Route choice set generation is important but not always essential. Recent work in two different directions obviates the need to explicitly generate choice sets. The first direction is the *sampling of alternatives* approach, which assumes a universal choice set from which a subset of routes is sampled with known probability distributions (Frejinger et al., 2009). Route choice models using such sampled choice sets are corrected using correction terms in the utility functions such that the parameter estimates become consistent to the universal choice set. The second direction is the link-based approach, where individuals' path choices are viewed as a series of successive link choices at each node, beginning from the trip origin (Fosgerau et al., 2013; Zimmerman et al., 2017). Specifically, recursive choice models are used within a dynamic programming framework to model the choice of a series of links. Most traditional (and still widely used) approaches to route choice modeling, however, relies on choice set generation.

the observed routes is called coverage. Many studies in the literature report coverage ranging from 22% to 96.6% for tolerance levels ranging from 0% to 30% for various route choice set generation algorithms (Bekhor et al., 2006; Hess et al., 2015; Prato and Bekhor, 2006, 2007; Rieser-Schüssler et al., 2013). Using this evaluation approach, coverage can be improved by generating more routes (which may increase the computation time), improving the algorithm itself, using a better algorithm, or combining choice sets from different algorithms. In doing so, however, one may end up with numerous irrelevant routes, which may not be considered by travelers and, therefore, potentially cause bias in estimation of choice model parameters and choice probabilities. In this context, a major drawback of the *trip-level* evaluation approach is that it does not offer a way to evaluate the generation of irrelevant routes, because the analyst cannot observe the travelers' consideration set from a single trip. Hence, most studies in the literature either do not address these irrelevant routes in the choice sets before route choice model estimation or filter out these routes based on various assumptions (see Rieser-Schüssler and Axhausen, 2009).

One way to overcome issues associated with *trip-level* evaluation is to perform the evaluation at an *OD pair level*, by comparing the *portfolio* of routes observed between an OD pair and those generated by the choice set generation algorithm between that OD pair. That is, if one can observe the routes of a sufficiently large number of trips between a given OD pair, one might get close to observing the travelers' consideration set for that OD pair. With increasing availability of large data sources (such as GPS data), it is now possible to observe a substantial number of trips made by multiple travelers between a given OD pair (see, for example, Lu et al., 2018). Therefore, using such data sources, analysts can compare observed choice sets with algorithm-generated choice sets at an *OD pair level* to evaluate the algorithm's ability to generate observed (i.e., relevant and/or considered) choice sets as well as the extent of generation of irrelevant routes. An evaluation of both aspects—the ability to generate relevant routes and the generation of irrelevant routes—can help improve choice set generation algorithms by increasing the capture of relevant routes while reducing irrelevant routes.

There are a few practical issues associated with evaluating choice set generation algorithms at an OD pair level. First, for any given OD pair, a sufficiently large number of trips should be observed for an unbiased evaluation of the choice set generation algorithms. Using a small number of observed trips is likely to cause biased evaluation because those trips might provide only a censored view of the travelers' consideration choice sets. The natural question is, how many trips

are necessary to observe the complete (or uncensored) consideration choice set between an OD pair? Conceptually, a rather large number of trips should be observed for each OD pair, but the data requirements may become prohibitively large to do so. Therefore, it may be pragmatic to determine a certain minimum number of trips that is, for practical purposes, sufficient to observe most of the consideration choice set.

The second practical issue is related to the spatial aggregation of trip ends (or OD locations). A disaggregate-level representation of OD locations for route choice analysis purposes is the link-level, where the OD pair is represented in the form of the network links at the trip ends; i.e., the first link of the route starting from the origin and the last link of the route ending at the destination. With such disaggregate spatial representation of OD locations, however, even with large data sources, it may not be easy to observe sufficient number of trips between link-level OD pairs. In addition, even if one observes a sufficient number of trips for a link-level OD pair, the observed route choices might not be diverse enough as these trips are typically made by only one or a few travelers (or, in case of freight travel, one or a few trucks). One way to overcome these issues is the consideration of spatially-aggregated OD pair locations, so it becomes easier to (1) observe sufficient number of trips for each (spatially) aggregated OD pair and (2) capture the diversity in route choices due to diversity in the travelers and their OD locations (or, in case of freight, diversity in the establishments trucks serve at the OD locations). Of course, spatial aggregation comes with its issues such as: (a) aggregation over large spatial units causing spurious diversity in route choices (due to the trip end locations being too far from each other) and (b) aggregation over observed choices of multiple travelers (or trucks) masking individual-level heterogeneity in choice sets. The key lies in choosing spatial units that are neither too large to cause spurious diversity nor too small to censor true diversity in route choices between an OD pair.[2] The question to be addressed here is, what is the optimal combination of the spatial aggregation and the minimum number of trips to observe for each OD pair?

[2] Although aggregation leads to homogeneous choice sets for different travelers between the same OD locations, it is not inconceivable that route alternatives chosen by one traveler are relevant to (and potentially considered by) another traveler. In fact, application of route choice models for prediction purposes in transport model systems with spatially-aggregated OD pairs potentially will benefit from allowing such aggregated choice sets that are inclusive of differences in traveler and spatial characteristics (Hoogendoorn-Lanser and Van Nes, 2004).

Another question is, can one use the insights from the above evaluation to improve choice set generation? In this context, a potentially effective approach that has not received much attention in the literature is to aggregate algorithm-generated choice sets over appropriately defined spatial units or OD pairs (similar to aggregating observed routes for evaluation purposes). Doing so can help in gaining the diversity needed in generated choice sets without having to generate too many routes for each disaggregate-level trip in the spatially aggregated OD pairs. A relevant question to be addressed here is, which is a better approach—generation of a large choice set at a disaggregate OD pair level or aggregation of small choice sets generated at a disaggregate OD pair level to a spatially-aggregated OD pair? Also, how can irrelevant route alternatives be reduced while increasing the capture of relevant alternatives in the choice set? Addressing these questions potentially can lead to substantial improvements to and/or effective use of existing choice set generation algorithms for route choice analysis.

1.1 Current Research

In view of the above discussion, this paper presents a new approach to evaluate truck route choice set generation algorithms, which is based on comparing the algorithm-generated choice sets between an origin-destination location pair (OD pair) against the portfolio of observed routes between that OD pair, using large streams of GPS data. The approach offers the ability to evaluate both the generation of relevant routes that are typically considered by the travelers and the generation of irrelevant (or extraneous) routes that are seldom chosen. Although the approach is general enough to be useful for evaluation of most route choice set generation algorithms (and for all modes of travel), this paper utilizes the proposed approach to evaluate the performance of the breath first search link elimination (BFS-LE) approach to generating route choice sets for freight-truck travel. The BFS-LE algorithm, proposed by Rieser-Schüssler et al. (2013), was chosen for evaluation as it has been gaining traction in the recent literature for generating route choice sets in high-resolution (i.e. with large number of nodes and links) transportation networks. Evaluation of other choice set generation algorithms is beyond the scope of this paper.

The study provides a carefully designed evaluation approach that is based on determining the optimal combination of (a) the spatial aggregation to represent trip OD locations and (b) the minimum number of trips to observe for each OD pair. Further, the evaluation uses metrics to

assess both the ability of route choice set generation algorithms to generate relevant routes (and the diversity therein) and the extent of generation of irrelevant (or extraneous) routes.

This study uses truck route choice data derived from large streams of truck GPS traces (more than 96 million truck GPS records) from more than 110,000 trucks traveling in Tampa, Florida. Given that the majority of route choice studies, other than a few exceptions (Arentze et al., 2012; Feng et al., 2013; Hess et al., 2015; Knorring et al., 2005), are in the context of passenger car or bicycle route choice, this study contributes to a currently small body of literature on generating route choice sets for modeling freight truck route choice.

It is worth noting that a recent study by Ton et al. (2018) makes use of a large amount of GPS data to observe multiple trips between an OD pair for bicycle route choice set generation. They evaluate whether the set of observed unique routes between an OD pair can be used to replace choice set generation altogether. The authors found that such data driven approach did not perform well (in predictions) with out-of-sample data. Unlike the Ton et al. (2018) study, this paper does not utilize the observed routes between an OD pair directly as a choice set for model estimation. Instead, the observed routes between an OD pair are used to evaluate the performance of an existing choice set generation algorithm and are treated as an asset to guide the analyst to make effective use of the algorithm for choice set generation.

In the rest of this paper, Section 2 describes the data used. Section 3 discusses the BFS-LE algorithm for route choice set generation, its implementation in this research, and the design of the evaluation approach. Section 4 presents the performance evaluation results and findings. Section 4 also presents estimation and validation results of route choice models estimated for different combinations of spatial aggregation and minimum number of trips observed per OD pair, and a comparison of route attributes of relevant routes and irrelevant routes. Section 5 summarizes and concludes the study.

2. Data

The primary data for this analysis, provided by the American Transportation Research Institute (ATRI), is truck-GPS data of more than 96 million GPS traces from a large fleet of trucks carrying GPS receivers (see Tahlyan et al., 2017). Geographically, the data spanned six counties of the Tampa Bay region in Florida—Hillsborough, Pinellas, Polk, Pasco, Hernando, and Citrus—and 15 miles beyond the six-county region. Temporally, the data were obtained for the first 15 days in

October 2015, December 2015, April 2016, and June 2016. The raw data were first converted into a database of truck trips using GPS-to-trip conversion algorithm developed by Thakur et al. (2015). The algorithm identifies trip ends by detecting potential stops (based on travel speed) of a certain minimum duration (five minutes) and using detailed land-use information to eliminate traffic stops and stops at rest areas. More than 1 million truck trips were generated along with the information on the OD location of each trip and other attributes such as trip start and end times and travel time. Subsequently, validation procedures were used to eliminate potentially problematic trips (due to GPS error or algorithmic error), highly circuitous trips with large detours potentially due to the algorithm missing a stop in between (detected by the ratio between direct OD distance and trip length less than 0.7), and trips less than five miles in length (as short truck trips would not have many route options). This resulted in more than 650,000 trips. It must be noted here that the five minute minimum duration to identify potential stops was determined after extensive review of the literature (see Thakur et al., 2015 and Kuppam et al., 2014 who also use the same dwell-time) and our own testing of various dwell-time values less than and more than 5 minutes. In this study, we observed that using larger (than 5 minutes) dwell time criteria increased the incidence of missing potential stops (false negatives) and smaller dwell times led to detecting stops on roadways which probably were traffic stops (false positives). This is perhaps because truck trip destination dwell times in urban areas are at least of 5 minutes duration (but not less than 5 minutes). Further, the criteria for using 0.7 as the ratio between direct OD distance and route length (to identify circuitous routes) was also determined using extensive testing with various values other than the 0.7 cutoff for the ratio (see Thakur et al., 2015).

For the trips generated above, the traveled routes were not necessarily readily-observable in the form of network links and nodes traversed between the OD locations. The raw GPS data of those trips had to be map-matched to the roadway network to derive the traveled routes. In this study, we used a high-resolution (i.e. with large number of nodes and links) NAVTEQ roadway network, available from the Florida Department of Transportation (FDOT), comprising more than 1.8 million links and 6.9 million nodes in the state.

To derive traveled routes for the truck trips generated from the GPS data, the GPS data were map-matched to the roadway network employing the procedures used in Kamali et al. (2016) and refined later by Tahlyan et al. (2017). High-frequency (i.e., closely spaced) GPS data are necessary for accurately deriving the traveled routes. GPS data for only about 50% of the derived

truck trips were spaced closely enough to avoid missing links in the routes derived from map-matching. For another 10% of the trips, some GPS data points could not be map-matched to the network, because the GPS data was not close to any link. After eliminating all such trips, traveled routes were derived for about 228,000 trips. For all these derived routes, an algorithm was developed and implemented to identify (and subsequently remove) routes with loops (or cycles) and routes that were too far from the original GPS data. Of the remaining 212,800 trips, 300 randomly selected routes were validated for consistency in the direction of travel, feasibility, and presence of large detours by evaluating the sequence of links and visualization on Google Earth. This exercise indicated 97% accuracy in the derived traveled routes as 9 out of 300 randomly selected routes had issues such as loops and large detours, primarily because one or few GPS points were mismatched to incorrect links on the network. The derived routes were considered as observed routes against which algorithm-generated route sets were evaluated. For more information on the data processing procedures and various assumptions used to detect trip ends and map-matching, readers are referred to Thakur et al. (2015) and Tahlyan et al. (2017).

For each of the derived trips, the corresponding route included information on the trip OD coordinates, corresponding traffic analysis zones (TAZs) defined in Florida's statewide travel demand model, and all the network links traversed by the truck between the OD locations. In addition, for each trip, several route attributes were computed, including route length, free flow travel times (from link-level speed limit information), travel costs (derived using the procedures by Torrey et al., 2014), number of intersections, left turns, right turns, and exit/entry ramps (each of these attributes was also computed per mile and per minute of travel), proportion of toll road length, and proportion of roads of several types (interstate highways, major arterials, minor arterials, collectors, local roads). In addition, to account for the similarity (or degree of overlap) of a route with other routes in the choice set for that same OD pair, a path-size attribute (Ben-Akiva and Bierlaire,1999) was computed as: $PS_i = \sum_{a \epsilon \Gamma_i} \left(\frac{l_a}{L_i}\right) \frac{1}{\sum_{j \epsilon C_n} \delta_{aj}}$, where Γ_i is the set of all links in path i between the OD pair n, l_a is the length of link a, L_i is the length of path i, C_n is the choice set of routes between the OD pair n, and δ_{aj} is equal to 1 if a route $j \epsilon C_n$ uses link a, 0 otherwise. The value of path-size for a route ranges between 0 and 1 (excluding zero), where a greater path-size value indicates smaller extent of overlap (and no overlap if path-size = 1).

3. Choice Set Generation and Evaluation Methodology

3.1 BFS-LE Algorithm and Its Implementation

The BFS-LE approach for route choice set generation belongs to the class of algorithms based on repeated least cost path search and is well-suited for extracting routes from large-scale, high-resolution networks. It is a link elimination approach (Azevedo et al., 1993) based on a repeated least cost path search, where links on the current shortest path are eliminated, one by one, to find subsequent least cost paths.[3] What distinguishes BFS-LE from other link elimination approaches is its use of a tree structure in which each node is a network. Beginning with the original network (which is the root node of the tree), any unique network obtained after the elimination of a link from a current least cost path is a node of the tree, as long as the network offers at least one feasible route for the OD pair under consideration. The nodes are arranged at various depths (d) in the tree based on the number of links eliminated. That is, $d = 1$ for a network obtained after removing any one link from the first least cost path between the OD pair in the root node (i.e., the original network), $d = 2$ for a network obtained after removing a link from the current least cost path between the OD pair in any of the nodes (or networks) at depth 1, and so on. For each node (network) at each depth, the links on the current shortest path between the OD pair under consideration comprise the breadth. The breadth first approach finishes the search for the next least cost path within a depth level, by removing links (one by one) on the current shortest paths in all nodes at that depth (i.e., across all breadths in that depth), before proceeding to the next depth level. The algorithm is aborted when a certain pre-defined number of routes are generated, a pre-defined time threshold is reached, or there are no more feasible routes to be found. The choice of the cost function to use (for least cost path search), the maximum number of routes to generate, and the time threshold are at the discretion of the analyst. For a more detailed description of the BFS-LE algorithm, readers are referred to Rieser-Schüssler et al. (2013).

To improve the computational performance of BFS-LE, Rieser-Schüssler et al. (2013) employ a topologically-equivalent network reduction technique in which nodes that are not

[3] Other variants of repeated least cost search algorithms are (1) simulation (Bierlaire and Frejinger, 2005; Prato and Bekhor, 2006; Ramming, 2001), where stochasticity in travelers' perceptions of travel costs and/or their preferences is simulated to generate multiple least cost routes, (2) path labeling (Ben-Akiva et al., 1984), where several least cost paths are obtained based on different criteria/labels for the cost function, and (3) link penalty (de la Barra et al., 1993), where links in the current shortest path are penalized with additional impedance to search for the next least cost path.

junctions of more than two links or dead-ends are eliminated and the corresponding links are merged to form a reduced (yet topologically equivalent) network for use in choice set generation. In addition, they use the A-star landmarks routing algorithm (Lefebvre and Balmer, 2007) instead of Dijkstra's algorithm (Dijkstra, 1959) for a quicker search of the least cost path.

It is worth noting that the BFS-LE is a deterministic choice set generation algorithm and, therefore, not a sampling (of alternatives) approach where the analyst uses probabilistically sampled alternatives to estimate route choice models (see, for example, Frejinger et al., 2009 for probabilistic sampling approaches for choice set generation).

In this study, the original network was coded and reduced to a topologically-equivalent network, and the BFS-LE algorithm was implemented in the Python programming language.[4] For the least cost path search, the free flow travel time was used as a cost function.[5] Following Dhakar and Srinivasan (2014), to avoid premature termination of the algorithm in situations with fewer than two outgoing links at the origin of a trip, the BFS-LE least cost search was started from the next node (in the route) that had at least two outgoing links. The BFS-LE generates routes that are different from each other even by one small network link. Since travelers may not consider routes with small deviations from each other as distinct, we considered a generated route to be a *unique* route (and, therefore, a part of the choice set) only if it is different from previously generated routes by at least 5%. Specifically, for a given OD pair, *unique* routes are determined (on the fly) using the commonality factor metric proposed by Cascetta et al. (1996), which determines the degree of similarity between two routes. Commonality factor (C_{ij}) between two routes i and j is: $C_{ij} = l_{ij}/\sqrt{L_i L_j}$, where l_{ij} is the length of shared portion between two routes and L_i and L_j are the lengths of the routes i and j, respectively. For a given OD pair, at every instance a route was generated from the BFS-LE algorithm, we considered it *unique* only if the commonality factors between that route and all previously generated unique routes were less than or equal to 0.95. We used 0.95 as the cutoff for commonality factors for it is neither too stringent (as a value of 1 would be) nor too lenient. Note here that a stringent value (closer to 1) would have led to generation of routes that are very similar to each other and a lenient value (like 0.85) would have led to generation of routes

[4] Python code written for implementing BFS-LE in this study is available here: https://github.com/dtahlyan/BFS_LE
[5] We explored the use of generalized travel time functions as well. However, our analysis showed better results with free flow travel times than those with the generalized functions explored for this study. Additional discussion on this is provided at the end of section 4.2.

that are significantly different but are still considered similar to each other. In addition, since the empirical context of this study is on freight truck route choice, to make the choice set generation more specific to freight trucks, we removed all those links from the network that were not accessible to trucks using the information available to us from the NAVTEQ network.

3.2 Evaluation Design

An important aspect of the evaluation in this paper was aimed at finding the appropriate combination of spatial aggregation (for trip ends) and minimum number of trips to be observed for each OD pair. These aspects are discussed first, followed by a discussion of the metrics used to evaluate how well the generated choice sets capture observed choice sets while not generating irrelevant routes that are not present in the observed choice sets.

3.2.1 Spatial Aggregation and Minimum Number of Trips to be Observed

Link-level aggregation: For all observed trips and their routes derived from the GPS data, the OD locations were represented in the form of network links at the trip ends; i.e., the first link of the route starting at the origin and the last link of the route ending at the destination. This is the most disaggregate representation of OD locations.

XY-level aggregation: The GPS locations of trip ends were aggregated by simply rounding off the longitude and latitude values from five decimal places to two decimal places. All trips with the OD coordinates matching up to the second decimal place were combined into a single XY-level OD pair. Such rounding leads to a spatial aggregation of roughly 1 km^2 at each of the trip ends. Rounding off to one or three decimal places was not adopted as it would have led to spatial aggregations which are too large (around 11 km^2) or too small (around 0.1 km^2), respectively. Note here that coarse spatial aggregation is not desirable as it leads to aggregation of trip ends that are far from each other into a single zone and very fine aggregation is not desirable as trip ends that are very close to each other end up in different zones.

TAZ level aggregation: The OD locations of observed trips were aggregated based on the TAZs defined in the Florida's Statewide Travel Demand Model, in which the state is divided into 5,403 TAZs. To avoid spurious diversity in the generated routes due to large-sized zones, we did not consider TAZ-level OD pairs with O/D TAZ sizes beyond 10 km^2. Further, we considered TAZ-level OD pairs with the following three levels of maximum TAZ size: 2 km^2, 5 km^2, and 10 km^2.

These numbers were chosen to understand and make an informed decision about the maximum TAZ size to work with.

Spatial clusters: Since large TAZs potentially cause spurious diversity in routes, spatial clustering was used to aggregate trip ends (or origin/destination location points of the trips) in larger (than 10 km^2) TAZs into smaller spatial clusters. After preliminary experimentation with different clustering techniques, the leader clustering technique (Hartigan, 1975) was used to divide the trip ends belonging to large TAZs into smaller clusters of radius 2 km while retaining the TAZ boundaries. An advantage of the leader clustering technique over the commonly used k-mean clustering technique is that the number of clusters is an output of the algorithm and need not be defined *a priori*.

Minimum number of trips to be observed: As discussed earlier, only OD pairs that have at least a minimum number of observed trips should be considered for a fair evaluation of choice set generation algorithms. To determine the minimum required number of trips, for each of the above-discussed aggregations, we considered OD pairs with the minimum number of observed trips of 20, 30, 50, and 100. This was done to understand the relationship between the number of observed trips and number of unique routes. Additional categories could have been tested but the current data did not have enough number of OD pairs with more than 100 observed trips. Also, it must be noted that the number here represents the trips per OD pair; not trips per user. Though having access to multiple trips per user would open avenues to gain a deeper understanding of user-level variability in route choice, it requires a significantly larger dataset than available to us. Further, since route choice analysis in practice is based on some spatial aggregation of the origin and destination locations (e.g., TAZs), aggregating the trips made by all trucks in the data that travelled between a given OD pair makes it closer to how the models are applied in practice.

Here, it must also be noted that while aggregating unique routes from link-level to larger spatial aggregations, the portion of the routes that were inside the corresponding spatial units were not removed. Doing so would have led to: (a) inaccuracies in calculating unique routes and overlaps, and (b) removal of the portions of the routes that were, in reality, used for travel.

3.2.2 Observed and Generated Unique Routes for Each Combination of Spatial Aggregation and Minimum No. of Trips

For each OD pair in each of the above categories, the observed routes of all trips (derived from the GPS data) were reduced to a set of unique routes using Cascetta et al.'s (1996) commonality factor

formula described earlier and applying an overlap threshold of 0.95. In addition to deriving the set of observed unique routes for each OD pair, the number of trips observed to have taken each unique route was also recorded.

Next, to generate route choice sets at different spatial resolutions, the BFS-LE algorithm was run to generate unique route choice sets at the link-level first with a limit of up to a maximum of 15 unique routes to be generated or for 1 hour, whichever was earlier, unless the algorithm stopped earlier due to completion of the tree. Such link-level generated choice sets were aggregated into larger spatial units discussed above using the commonality factor formula with an overlap threshold of 0.95.

3.2.3 Evaluation Metrics

Let the set of observed unique routes for an OD pair n be $O_n = \{o_1, o_2, \dots, o_i, \dots, o_{I_n}\}$ and the set of generated unique routes for that OD pair be $G_n = \{g_1, g_2, \dots, g_j, \dots, g_{J_n}\}$, where i is the index for an observed unique route, j is the index for a generated unique route, I_n and J_n are the number of observed unique routes and generated unique routes, respectively, in the n^{th} OD pair. Let k_i be the number of trips observed to have taken the unique route i. To measure the performance of BFS-LE choice set generation implemented in this study, we devised three metrics to compare the observed and generated unique route sets at an OD pair level—(1) false negative error, (2) weighted false negative error, and (3) false positive error—each of which is discussed next.

False negative error (ε_n^-): False negative error for an OD pair n is the proportion of observed unique routes that are not generated by the choice set generation algorithm. Mathematically, $\varepsilon_n^- = 1 - \frac{\sum_{i=1}^{I_n} \delta_i}{I_n}$, where $\delta_i = 1$ if the commonality factor C_{ij} between the observed unique route i and any of the generated unique routes $j \in G_n$ is greater than 0.95, zero otherwise. ε_n^- ranges between 0 and 1; the most desirable value is 0 (when all observed routes are generated) and least desirable value is 1 (when none of the observed routes is generated).

Weighted false negative error (ε_{wn}^-): Weighted false negative error is the proportion of observed trips (not unique routes) whose observed unique routes are not generated by the choice set generation algorithm. It is a weighted version of the false negative error, where the capture (by the choice set generation algorithm) of each observed unique route is weighted by the proportion of trips taking that route. Specifically, $\varepsilon_{wn}^- = 1 - \frac{\sum_{i=1}^{I_n} k_i \delta_i}{\sum_{i=1}^{I_n} k_i}$. The unweighted metric (ε_n^-) equally

penalizes the choice set generation algorithm for not capturing any observed unique route, regardless of the usage of that route. The weighted metric overcomes this shortcoming by penalizing an uncaptured route based on the extent of its usage.

False positive error (ε_n^+): False positive error for an OD pair n is the proportion of generated unique routes that are not presented in the observed unique routes set. This metric provides a measure of the irrelevant (or extraneous) routes generated that are not observed to have been chosen by the traveler. Specifically, $\varepsilon_n^+ = 1 - \frac{\sum_{j=1}^{J_n} \delta_j}{J_n}$, where $\delta_j = 1$ if the commonality factor C_{ji} between the generated unique route j and any of the observed unique routes $i \in O_n$ is greater than 0.95, zero otherwise. ε_n^+ ranges between 0 and 1; the most desirable value is 0 (when all generated routes are observed) and least desirable value is 1 (when none of the generated routes are observed). As discussed earlier, a trip-level evaluation of the choice set generation algorithms does not allow one to evaluate false positives (i.e., the generation of extraneous routes).

3.2.4 Performance Evaluation

First, to evaluate the performance of the implemented BFS-LE approach, the above discussed error metrics were compared at various levels of spatial aggregation and minimum number of trips per OD pair. The same metrics were used to determine appropriate combination of spatial aggregation and minimum number of trips for the performance evaluation. Second, for various spatial aggregations ranging from link-level to TAZ-level, we recomputed the error metrics for generated choice sets constructed out of implementing BFS-LE with the following limits on the maximum number of routes generated for each link-level OD pair: 5, 10, 15, 20, and no limit. The time limit to abort the algorithm was set to 1 hour in all cases. The resulting error metrics were analyzed to determine which is a better approach – generation of a large choice set at a disaggregate OD pair level or aggregation of small choice sets generated at a disaggregate OD pair level to a spatially aggregated OD pair? To further examine this, route choice models were estimated and applied (on validation data) using choice sets constructed at link-level and TAZ-level aggregations; constructed from a maximum of 5 and 15 BFS-LE routes generated at the link-level. Finally, various attributes of routes that were observed as well as algorithm-generated (i.e. relevant routes) were compared with those of the extraneous (or irrelevant) routes that were generated but not observed.

4. Evaluation Results

4.1 OD Pair-level Evaluation of Choice Set Generation Algorithm at Different Combinations of Spatial Aggregation and Minimum Number of Observed Trips

Table 1 presents the evaluation results for each combination of spatial aggregation and minimum number of observed trips considered at an OD pair level. This table represents a total of 82,738 truck trips extracted from the 212,800 trips for which routes were derived (and validated). While the 212,800 trips belong to 53,009 unique TAZ-level OD pairs defined in the state of Florida's travel demand model (and 130,090 link-level OD pairs), these 82,738 trips belong to only 2,125 TAZ-level OD pairs (and 23,112 link-level OD pairs). The 82,738 trips were aggregated to different spatial levels, while considering the minimum number of trips available for each spatially aggregated OD pair. Specifically, only trips which belonged to an OD pair with at least 20 observed routes (in at least one of the spatial aggregations) were retained. The remaining trips were discarded as those trips belonged to OD pairs where the observed trips are not enough to observe an unbiased route choice set. Various observations and inferences from this table are discussed next.

4.1.1 Observed and Generated OD Pair-level Choice Sets

First, the columns titled 'No. of OD Pairs' and 'No. of Trips' present the observed data available for each combination of spatial aggregation and minimum number of observed trips. For example, at least 20 trips were observed for 615 OD pairs at the link-level, and 29,003 trips were observed between these 615 OD pairs. As expected, for a given spatial aggregation, the number of OD pairs with available data decreased as the minimum number of trips increased from 20 to 100. Likewise, for a given minimum number of trips, the number of OD pairs with available data increased from a finer spatial resolution to a higher spatial aggregation.

Second, one can infer from the column titled 'No. of Observed Unique Routes' that the average number of observed unique routes per OD pair increased with increase in spatial aggregation and/or with increase in the minimum number of trips observed. In the context of spatial aggregation, a visual inspection of trip ends in different OD pairs suggested that increasing the TAZ size beyond 2 km^2 led to a spurious increase in unique routes due to the trip ends within a TAZ becoming too far from each other. In the context of the role of minimum number of trips observed, the number of unique routes observed did not stabilize even after observing a minimum of 50 trips per OD pair. However, it can be observed that the increase of the average number of

observed unique routes with respect to the minimum number of observed trips occurs at a decreasing rate, with the lowest increase in the number of additional observed unique routes per unit increase in the minimum number of trips observed occurring between 50 to 100 minimum trips per OD pair. This non-linear relationship between the number of observed unique routes and number of observed trips between an OD pair is also reported in Luong et al. (2018), where number of observed unique routes in an OD pair is modeled as a function of observed trips, OD, and network characteristics using a truncated negative binomial regression. Besides, there were some outlier OD pairs (with high number of observed unique routes) that skewed the reported average values in Table 1 for OD pairs with a minimum of 100 trips. Therefore, for pragmatic reasons (such as not to lose substantial amount of data), we determined that observing a minimum of 50 trips per OD pair was sufficient to derive an observed route choice set for evaluation purposes.

Third, it can be observed from comparing the column titled 'No. of Generated Unique Routes' to the preceding column that the number of generated routes was generally greater than the number of observed routes for an OD pair. Further, as expected, the number of generated unique routes increased with increase in spatial aggregation, but at a higher rate than the increase in the number of observed unique routes. This suggests the possibility of increase in the generation of irrelevant routes with increasing spatial aggregation.

4.1.2 OD Pair-level Error Rates

Several observations can be made from the error metrics, which are reported in the last three columns in Table 1. First, the weighted false negative errors, ranging from 11% to 26%, were smaller than their unweighted counter parts, which range from 34% to 55%. As discussed earlier, the unweighted metric does not take into consideration the extent of usage of a route; whereas the weighted metric computes the errors based on usage of routes, with the errors on more (less) used routes carrying a greater (lower) weightage. In fact, the average weighted false negative errors were under 20% for most combinations of spatial aggregation and minimum number of observed trips. Therefore, one can infer that the BFS-LE performs well in capturing the more frequently-used routes than the less frequently used routes in the dataset.

Second, for any given minimum number of trips between an OD pair, the weighted false negative errors were lowest at a spatial aggregation of TAZs of up to 2 km^2. This suggests that choice sets created by aggregating the generated routes from a spatial resolution of TAZs of up to 2 km^2 can help in improving the capture of observed routes. Interestingly, the improvement in

weighted false negative errors was lost when larger-sized TAZs were included, perhaps because the observed routes between larger TAZs would have spurious diversity due to the trip ends being too far from each other. Also, the error rates for spatial aggregations of XY-level and spatial clusters were higher than those of smaller-sized TAZs. This is likely because TAZs are typically created keeping in view the transportation network structure around (as opposed to the other aggregations we created) and that small-sized TAZs provided an optimal mix of diversity in trip-starting and trip-ending links (which results in diverse routes between the TAZs), while keeping the trip ends within a concentrated area to avoid spurious diversity. It is also interesting to note that the standard deviations of weighted negative errors were smallest for the spatial aggregation of TAZ-level of up to 2 km^2. All these results suggest that route choice sets created out of aggregating routes generated between different trip-end links of small-sized TAZ pairs can potentially capture a large share of observed routes.

Third, as can be observed from the column titled 'False Positive Error', the proportion of extraneous/irrelevant routes in the generated choice sets increased from the link-level to any other spatial aggregation considered in this study. As expected, increasing the capture of relevant routes (i.e., decreasing weighted false negative error rates) through spatial aggregation comes with an increase in extraneous routes as well. Interestingly, however, the average false positive error rates were not very different across different spatial aggregations other than the link-level. It should be noted here that the overall range of false positive errors across various aggregations is high. Specifically, on an average, at least 80% of the generated unique routes were not observed. This might be in part because the diversity in truck routes tends to be smaller than that in the generated routes. Anyhow, the high rate of false positive errors (or irrelevant routes) could have adverse implications for route choice predictions (see Bliemer and Bovy, 2008). Note, however, that previous studies have not reported false positive errors, which makes it difficult to benchmark what is an acceptable rate for false positive errors. This warrants a need for additional empirical evidence on false positive error rates from other empirical contexts and choice set generation procedures. In fact, our results raise the question if choice sets generated from other algorithms will have similarly high generation of irrelevant routes; a topic of relevance for further research. Further, as pointed out by a reviewer, the issue of false negatives might be more a problem when route choice models are used in stochastic traffic assignment (where some flow is always assigned to all routes in the choice set) than in deterministic traffic assignment where not all routes in the

choice set are assigned flows. And the extent of the problem due to false positive errors in the generated choice sets depends on the level of differences between observed routes and generated false positive routes.

Overall, the above-discussed results suggest the potential benefits of OD pair-level evaluation of choice set generation algorithms over the traditionally used trip-level evaluation. As importantly, aggregating the generated choice sets over carefully defined spatial units (which happens to be TAZs of up to 2 km^2 in this empirical context) can help improve the capture of relevant routes for subsequent route choice modeling and prediction. However, it should be noted that spatial aggregation also results in an increase in the number of irrelevant routes.

Additional analysis was carried out in the following directions: (a) comparison of OD pair-level evaluations results to trip-level evaluation results, and (b) evaluation of BFS-LE choice sets for different thresholds of overlap between observed and generated routes. For interested readers, results and the corresponding discussion from the additional analysis are presented in an Appendix.

4.2 Which is Better: Spatial Aggregation of a Limited Number of Generated Routes or Increasing the Number of Routes Generated from BFS-LE?

Findings from Table 1 suggested that spatial aggregation of generated routes can potentially help in increasing the capture of observed routes. Now, we examine if one can increase the capture of observed routes by generating a small number of routes at the link-level OD pairs and then spatially aggregating them to TAZ-level (instead of generating large number of routes at the link level). The hypothesis is that generating a smaller number of unique routes at the link-level and aggregating them spatially will lead to sufficient diversity in the generated choice sets. By doing so, one can reduce the computational burden of generating a large number of routes at the disaggregate level.

Table 2 presents error measures for choice sets generated from different limits on the maximum number of generated unique routes at the link-level—5, 10, 15, 20, and no limit—for two different spatial aggregations—TAZ-level (of up to 2 km^2 size) and link-level. Note that the average weighted false negative error values (and the standard deviations), for both the TAZ-level and link-level aggregations, did not vary from choice sets constructed from a maximum of 5 unique BFS-LE routes to those generated from a maximum of 20 or more. This indicates that the BFS-LE is able to do its best (in generating observed routes) early on in its implementation. Running the algorithm longer to generate more than 5 unique routes did not improve the performance of BFS-

LE. In fact, false positive errors increased with the increase in the number of generated routes beyond 5. What helped more than generating beyond 5 BFS-LE unique routes is aggregation (from link-level to TAZ level). That is, aggregating (from link-level to TAZ level) up to only 5 BFS-LE unique routes helped in a better capture of observed routes than generating a larger number (up to 20 or more) of routes without aggregation. This may be because aggregation helps bring more diversity in the generated routes than running BFS-LE for longer. This is perhaps the first study to demonstrate that using choice set generation algorithms to generate more and more routes might not help as much as aggregating a small number of algorithm-generated routes.

The above results and discussion suggest that a computationally efficient way to increase the diversity of generated choice sets (and thereby increase the coverage of observed routes) is to aggregate a limited number of link-level choice sets generated from close by locations. Of course, as can be observed from the table, false positive errors increase with spatial aggregation. Therefore, one must estimate and apply route choice models (and compare prediction ability) using different choice sets to evaluate if the benefits due to decrease in false negative errors outweigh dis-benefits due to increase in false positive errors.

Note that the above findings are based on route choice sets generated using free flow travel times. Following a reviewer's suggestion to examine the robustness of our findings with respect to the choice of travel time function used to generate route choice sets, we conducted additional analysis with a generalized travel time function used to generate route choice sets. Following Biswas et al. (2019), we used a statistical model (based on observed travel time values from the GPS data) to devise a generalized travel time function that established a relationship between route characteristics (link road type and toll information) and observed travel time values, which was subsequently used to impute congested travel time values on each link in the network. These congested travel time values were then used to generate route choice sets and compute the weighted false negative and false positive error values. The results obtained using the generalized travel time function showed similar trends (as in the case of free flow travel time function) where the spatially aggregated choice sets had lower weighted false negative errors. Further, we also found that the values of error metrics reported in Table 2 using free flow travel time function were lower than those obtained using the generalized travel time function. This is possibly because a majority of observed routes in our data overlapped significantly with the corresponding shortest free flow time route (as seen in our additional analysis presented in the Appendix). Therefore, we

retained the choice sets generated using the free flow travel time function for all the analysis in this paper.

4.3 Estimation and Validation of Route Choice Models with Different Choice Sets

To further evaluate the hypothesis that aggregating a limited number of BFS-LE routes leads to better choice sets than generating a large number of routes from the BFS-LE without aggregation, we estimated and applied a series of route choice models from choice sets at link-level and TAZ-level aggregations constructed from up to a maximum of 5 or 15 BFS-LE alternatives. All the models were estimated on a sample of 6,453 trips and were applied on a validation sample of 1,758 trips (~20 % of the total sample) to evaluate the impact of choice set composition on route choice prediction.[6]

Three different empirical specifications were used: (a) path size logit (Ben-Akiva and Bierlaire, 1999), (b) error components logit (Frejinger and Bierlaire, 2007) and (b) error components logit with random coefficients on route attributes. The path size logit (PSL) model structure employs the theory of aggregation of alternatives (see Ben-Akiva and Lerman, 1985) to recognize that a route that overlaps with another may not be perceived as a distinct alternative. To do so, the utility of a route is corrected by including natural logarithm of a path size (PS) attribute. The utility associated with a route i for observation n is written as $U_{in} = \beta' X_{in} + \beta_{PS} lnPS_{in} + \varepsilon_{in}$, where X_{in} is a vector of observed attributes of route i, β is a corresponding vector of parameters, PS_{in} is the path size attribute for route i, β_{PS} is a parameter corresponding to the path size variable, and ε_{in} is the random utility component assumed to be to IID Gumbel distributed. The probability (P_{in}) of choosing a route i by a truck n facing a choice set C_n is:

$$P_{in} = \frac{\exp(\beta' X_{in} + \beta_{PS} lnPS_{in})}{\sum_{j \epsilon C_n} \exp(\beta' X_{jn} + \beta_{PS} lnPS_{jn})} \quad (1)$$

[6] A total of 8,211 trips were available from TAZ-level OD pairs that had a minimum of 50 trips per OD pair (Table 1 reports this number). Of these 8,211 trips, we used 6,453 (80%) for route choice model estimation and 1,758 (20%) for model validation. Although it would be desirable to use all available data for model estimation, only these trips were used in this analysis to maintain consistency among the observations used for model estimations across various spatial aggregations. That is, the same set of trips ought to be used to evaluate the route choice sets generated at TAZ-level as well as link-level so that any differences in model performance may be attributed to differences in spatial aggregation as opposed to differences in the data used.

The path size logit formulation accommodates correlations between route alternatives due to physical overlap between routes. However, correlations between route alternatives might also arise due to unobserved factors that are not attributable to physical overlap. To capture such correlations, we use the error components logit (ECL) model structure proposed by Frejinger and Bierlaire (2007) for route choice models. Specifically, the ECL structure helps capture the perceptual correlations among route alternatives passing through a same labelled road without necessarily overlapping. As an illustration, according to the ECL model structure, the utilities of three routes i, j and k in a choice situation faced by a truck in observation n are written as:

$$U_{in} = \beta' X_{in} + \beta_{PS} lnPS_{in} + \sigma_a \sqrt{L_{in,a}} \xi_{n_a} + \sigma_b \sqrt{L_{in,b}} \xi_{n_b} + \varepsilon_{in} \quad (2)$$

$$U_{jn} = \beta' X_{jn} + \beta_{PS} lnPS_{jn} + \sigma_a \sqrt{L_{jn,a}} \xi_{n_a} + \varepsilon_{jn} \quad (3)$$

$$U_{kn} = \beta' X_{kn} + \beta_{PS} lnPS_{kn} + \sigma_a \sqrt{L_{kn,a}} \xi_{n_a} + \sigma_b \sqrt{L_{kn,b}} \xi_{n_b} + \varepsilon_{kn} \quad (4)$$

where $L_{in,a}$, $L_{jn,a}$, and $L_{kn,a}$ are the distances covered by routes i, j, and k, respectively, on the named road/label a. Similarly, $L_{in,b}$, and $L_{kn,b}$ are the distances covered by routes i, and k, respectively, on the named road/label b. Further, ξ_{n_a} and ξ_{n_b} are random variables, assumed to be standard normal and distributed independently and identically across observations. In addition to such ECL models, to account for unobserved heterogeneity in sensitivity to route attributes, we allowed random (normally distributed) parameters for the coefficients of route characteristics.

The PSL model estimation was carried out using the maximum likelihood estimation technique. The ECL and ECL with random parameters models were estimated using the maximum simulated likelihood estimation approach where 400 Halton draws (Bhat, 2003) were used to evaluate the multi-dimensional integral of the likelihood function. The choice sets used for all model estimations were augmented with the chosen routes (if the chosen routes were not already generated).

4.3.1 Estimation Results of Route Choice Models

Table 3 presents estimation results of the ECL model with a random coefficient on the travel time variable – estimated with TAZ-level choice sets aggregated from up to 5 BFS-LE unique routes at the link level. The parameter estimates suggest that routes with a lower travel time, lower travel cost, smaller proportion (in length) of tolled routes, smaller number of turns and ramps per minute,

and those with a higher proportion of road length on major highways were preferred over other routes. However, there is significant unobserved heterogeneity in the sensitivity to travel time, as evident by the random coefficient on the travel time variable. Further, error components corresponding to the following four named roads turned out to be statistically significant: Interstate 4 (I-4), Interstate 75 (I-75), Polk parkway (also called Florida state road 570), and United States Route 19 (US-19).

The average value of travel time from the parameter estimates reported in Table 3 is $46.15/hour and the standard deviation is $287.12/hour. Such a wide distribution of value of time is akin to the wide range of values of time reported by Shams et al. (2017) using a survey conducted among freight carriers, shippers and forwarders in the same geographical context (Florida). However, since the travel time coefficient in our random parameters model reported in Table 3 is assumed to be normally distributed, the percentage of negative values of time is rather high (above 43%). This could be attributed to multiple reasons, including: (a) the normal distributional assumption used for travel time coefficient is not appropriate, (b) travel costs are highly correlated with the travel times and do not include tolls (percentage of tolled roads are included as a separate variable), or (c) because not all trips take the shortest free flow time path. Addressing these issues is an avenue for future research. Therefore, for the sake of robust findings, we estimated and evaluated several other models, which do not include a random coefficient on travel time, as discussed next. [7]

Including the model reported in Table 3, a total of 16 models were estimated whose estimation results are not reported here to conserve space. The parameter interpretation was consistent across most models with some models having counter intuitive interpretation to the travel time variable. Specifically, the parameter estimates for the travel time and travel cost variables in a few models had counter intuitive sign, likely due to high correlation between travel time and travel cost in the corresponding datasets. Table 4 reports the following model fit measures on the estimation data for 15 of these models: log-likelihood value at convergence ($\mathcal{LL}_C$); log-likelihood value for equal shares model ($\mathcal{LL}_{ES}$); adjusted rho-square ($\overline{\rho^2}$), calculated as

[7] Such issues with estimation of value of time are not uncommon, particularly when the cost variable is correlated with the time variable making it difficult to estimate positive values of travel time savings. However, as discussed later, since the findings in the context of choice set generation are similar across all 16 route choice models estimated in this study, the findings are robust to model specification issues.

$\left(1 - \frac{\mathcal{LL}_C - k}{\mathcal{LL}_{ES}}\right)$; Akaike information criterion (AIC), calculated as $[2k - 2\ln(LL_C)]$; and Bayesian information criterion (BIC), calculated as $[\ln(n)\,k - 2\ln(LL_c)]$. Here, n is the number of observations in the data (which is 6,453 for all models) and k is the number of estimated parameters. For any given choice set, as expected, models with error components and random coefficients show better fit to estimation data than simpler models. For example, for the choice set at link level aggregation with up to 5 BFS-LE alternatives, the path size logit model shows inferior fit than advanced models, as seen from lower $\overline{\rho^2}$ value and higher $\mathcal{LL}_C$, AIC, BIC values for the path size model. These results align with expectations and support the results reported by other studies (for example, Frejinger and Bierlaire, 2007). Of course, one should not use data fit measures for comparing the performance of models with different choice sets. Therefore, the next sub-section compares measures of route choice predictions using different choice sets.

4.3.2 Validation Results with Route Choice Models

As indicated earlier, a validation sample of 1,758 trips was used to evaluate the impact of choice set composition on route choice prediction. For all these trips, it is important to note that the choice sets used for prediction included the chosen route only if it was generated (so that the prediction results can be used to evaluate the generated choice sets). The number of cases for which the chosen route was not generated were 303 and 223 for link-level choice sets of up to 5 and 15 BFS-LE routes, respectively. And the number of cases for which the chosen route was not generated were 183 for both the choice sets at TAZ-level aggregation. The metric used for validation of route choice predictions (on the validation dataset) is based on expected overlap of route choice predictions with the observed route.[8] Specifically, for a trip (or observation) n with route choice set $\{1, \dots, 2, \dots, i, \dots, I\}$ and chosen route r, the expected overlap was $E(O)_n = \sum_{i=1}^{I} p_i C_{ir}$, where p_i is the probability of choosing route i from the choice set and C_{ir} is the proportion of route i common with the chosen route r. The average value (and standard deviation) of expected overlap across all trips in the validation data was used to evaluate route choice predictions by different models.

[8] Another measure of predictive ability would be the log-likelihood value over the validation dataset. However, such predictive log-likelihood values cannot be calculated when the chosen alternative is not in the choice set as it was not generated by the choice set generation algorithm (to be precise, the predictive likelihood s zero and log-likelihood is negative infinity). Therefore, the expected overlap measure was chosen as a metric of predictive performance.

Table 5 reports the validation results. Interestingly, the results suggest that the models estimated with link-level choice sets of up to 5 or 15 BFS-LE generated routes have, on average, better expected overlap (hence, better predictive ability) than the models estimated with choice sets at TAZ-level aggregations. This pattern holds for all model specifications – path size logit, ECL and ECL with random coefficients. A possible explanation for this result is that choice sets aggregated to the TAZ-level have a greater number of irrelevant (or extraneous) routes than link-level choice sets. As discussed in Section 4.2, spatial aggregation of choice sets increases the diversity of generated routes and thereby improves the coverage of relevant routes. At the same time, spatial aggregation increases the presence of irrelevant routes whose overlap with the chosen route is much smaller than that of relevant routes. This is likely a reason for a lower value of average expected overlap for TAZ-level choice sets than those for link-level choice sets. Although not in favor of the proposed spatial aggregation approach to building route choice sets, this finding is not totally unexpected; such adverse effects of irrelevant routes on the prediction capability of route choice models has been pointed out by other studies as well (Bliemer and Bovy, 2008). The results also suggest that the prediction benefits of spatial aggregation approach to choose set construction can potentially be harnessed better if irrelevant routes are eliminated from the aggregated choice sets.

Another interesting result is that advanced model structures, such as ECL and ECL with random parameters exhibit inferior prediction capabilities (as measured by average expected overlap) when compared to the simpler, path size logit structure. This is in contrast with the model fit trends discussed earlier, where advanced model structures exhibited better fit to estimation data. In addition, this finding appears to contrast with the results of Frejinger and Bierlaire (2007) who demonstrated better predictive loglikelihood values for ECL models over path size logit models. Specifically, advanced models like ECL in our case had lower mean expected overlap than the PSL models, while Frejinger and Bierlaire (2007) report a higher value of predictive likelihood for ECL models when compared to PSL models. It is worth noting, however, given our focus on the role of choice set composition in predictions, that we did not include the chosen alternative in the choice set used for prediction unless it was generated by the BFS-LE. However, choice sets in the Frejinger and Bierlaire (2007) paper include the observed route regardless of whether it was generated or not. Therefore, it is our conjecture that prediction abilities of different route choice model structures might depend considerably on the choice set composition. Further research is

necessary to understand the role of choice set composition on the predictive performance of advanced model structures.

4.4 Comparison of the Characteristics of Observed and Generated Choice Sets

Table 6 presents a comparison of characteristics of the routes that were observed as well as generated (i.e., relevant routes captured in generated choice sets) to routes that were generated but not observed (i.e., extraneous routes). This comparison suggested that extraneous routes in the generated choice sets were generally longer, had a greater proportion of tolled roads, involved a greater proportion of the route through smaller roads (such as minor arterials, collectors, and local roads), more network links per mile, and more intersections and turns than the relevant routes in the generated choice sets. A visual examination of several extraneous routes suggested that many such routes involve getting off an interstate highway to smaller roads and then getting back on to the interstate highway.

A potential use of the comparison presented above is in devising strategies to remove extraneous routes from the TAZ-level choice sets in a post-processing step. For example, further analysis may be conducted to identify thresholds (either deterministic or probabilistic) on selected route-level attributes such as maximum number of turns/intersections per mile. Once such thresholds are identified, generated routes that do not meet the threshold criteria may be eliminated from the choice set. Another approach is to device a probabilistic approach that corrects route choice probabilities based on how likely a route is to be extraneous. Exploration of such strategies is an avenue for future research.

Note that the BFS-LE algorithm evaluated in this study is a is a link-elimination algorithm that operates on link-level characteristics, not on route-level characteristics, which makes it difficult to use route-level characteristics in this algorithm. Therefore, route-level characteristics – such as proportion of routes through smaller roads, no. of links per mile, number of intersections and turns, and any freight/truck-specific route-level characteristics – can be considered in a post-processing step to refine the choice sets generated from the algorithm. Since the number of BFS-LE generated routes is typically small (compared to all possible routes available between an OD pair), it is easier to work with route-level characteristics in the post-processing stage as opposed to considering computationally expensive procedures to consider route-level attributes at the choice set generation stage.

5. Summary and Conclusions

This study proposed a new approach to evaluate truck route choice set generation algorithms and derived guidance on using the algorithms for effective generation of choice sets for modeling truck route choice. Specifically, route choice sets generated from the breadth first search link elimination (BFS-LE) algorithm were evaluated against observed truck routes derived from large streams of GPS traces of a sizeable truck fleet in the Tampa Bay region of Florida. A carefully-designed evaluation approach was presented to arrive at an appropriate combination of spatial aggregation and minimum number of trips to be observed between each OD location for evaluating algorithm-generated route choice sets. The evaluation assesses both the ability to generate relevant routes that are considered by travelers and the generation of irrelevant (or extraneous) routes that are seldom chosen. Based on the evaluation, the study offered guidance on effectively using the BFS-LE approach to maximize the generation of relevant truck routes. Further, route choice models were estimated and applied on validation datasets to confirm findings from the above evaluation. Lastly, a comparison of route attributes of relevant and irrelevant routes was done to understand systematic differences in route characteristics of the relevant and irrelevant routes.

The results demonstrate the benefit of evaluating algorithm-generated choice sets against observed choice sets from large datasets at a spatially aggregated OD-pair level (instead of performing trip-level evaluations). Doing so helps in evaluating the ability to generate relevant routes as well as the generation of irrelevant routes. Based on the evaluation results, it was found that a carefully chosen spatial aggregation (of generated routes) can help improve the coverage of relevant routes while also reducing the need to generate substantial number of routes for each trip. The implication is that an effective and computationally effective use of the BFS-LE algorithm for generating truck route choice sets is to generate a small number of routes at a disaggregate OD pair level and then aggregate such routes from nearby OD locations.

This is perhaps the first study to demonstrate that using choice set generation algorithms to generate a large number of routes might not help as much as spatially aggregating a small number of algorithm-generated routes. In the empirical context considered in this study (truck routes in Florida), generating up to 5 unique routes per trip and aggregating such BFS-LE routes across the trips ending within traffic analysis zones of up to 2 square kilometers resulted in smaller errors than generating up to 20 unique BFS-LE routes for each trip. This is perhaps because spatial

aggregation helps bring more heterogeneity/diversity in the routes than running BFS-LE for longer.

The spatial aggregation approach, however, is not without its disadvantages. A greater presence of irrelevant routes in spatially aggregated route choice sets might offset (or even outdo) the benefits of increased coverage of relevant routes in the context of route choice prediction. For these reasons, our empirical results with data from Florida showed an inferior predictive ability of route choice models with spatially aggregated choice sets than those with disaggregate choice sets. Therefore, the prediction benefits of spatial aggregation approach to choice set construction (which helps increase the coverage of relevant alternatives) can potentially be better harnessed by eliminating irrelevant routes from the choice sets. Exploration of alternative ways to eliminate irrelevant routes is a fruitful avenue for future research.

The findings of this study also suggest that extraneous routes generated by the BFS-LE are generally longer, have a greater proportion of tolled roads, and involve a greater proportion of the route through smaller roads (such as minor arterials, collectors, and local roads), more network links per mile, and more intersections and turns than observed truck routes in Florida. Using such results, future research can focus on the development of approaches to either eliminate or reduce the probability of extraneous routes from aggregated route choice sets. The advantage of the post-processing approach is that the algorithms used to generate choice sets need not consider route-level characteristics, because the post-processing step can do so. The BFS-LE algorithm, for example, operates on link-level characteristics which makes it computationally less expensive when compared to other algorithms in the literature that need route-level characteristics to generate choice sets. Since the number of BFS-LE generated routes is typically small (compared to all possible routes available between an OD pair), it is easier to work with route-level characteristics in the post-processing stage as opposed to doing so at the choice set generation stage.

Another interesting finding of this study is in the context of prediction ability of different route choice model structures. When the generated choice sets did not include the chosen route, advanced choice modelling structures (such as random coefficients and/or error components models) showed inferior prediction performance when compared to a simpler, pathsize logit model structure. The influence of choice set composition on the prediction performance of different route choice model structures is an important avenue for subsequent research.

In future research, it would be useful to gain a better understanding of the data requirements for extending the analysis done in this study. Specifically, our study made use of temporally spread out GPS data (15 days of data from each of 4 different months) to obtains trips in various OD pairs. It would be good to determine for how long trips in an OD pair must be observed to obtain an unbiased observed route choice set. Are a minimum 50 trips observed over a period of 15 days enough or do we need to look for trips which are temporally more spread out to better capture the observed choice sets? Another avenue for future research is to formulate and utilize link-level error metrics to evaluate choice set generation (as opposed to route-level error metrics used in this study). Specifically, it would be useful to do an analysis of percentage of links which are present in the observed route choice set are generated by the choice set generation algorithm and the percentage of links which are generated by the choice set generation algorithm but not present in the observed routes.

Acknowledgements

This study was partially funded by the Florida Department of Transportation and the United States Department of Transportation through the University Transportation Center (UTC) called Center for Advanced Infrastructure and Transportation (CAIT). Support received to the first author through the Walter P. Murphy Fellowship at the Northwestern University and partial funding to the second author from a faculty startup grant by the Indian Institute of Science (IISc) is also acknowledged. The authors are grateful to American Transportation Research Institute (ATRI) for providing the GPS data. Thanks are due to Siva Srinivasan for helpful discussions on implementation aspects of the BFS-LE algorithm. An early version of this paper was presented at the 2018 Transportation Research Board Annual Meeting. Four anonymous reviewers and the Editor provided valuable comments that helped improved the overall study and its positioning.

References

1. Arentze, T., Feng, T., Timmermans, H., Robroeks, J., 2012. Context-dependent influence of road attributes and pricing policies on route choice behavior of truck drivers: results of a conjoint choice experiment. *Transportation* 39 (6), 1173-1188.
2. Azevedo, J., Costa, M.E.O.S., Madeira, J.J.E.S., Martins, E.Q.V., 1993. An algorithm for the ranking of shortest paths. *European Journal of Operational Research* 69(1), 97-106.

3. Bekhor, S., Ben-Akiva, M.E., Ramming, M.S., 2006. Evaluation of choice set generation algorithms for route choice models. *Annals of Operations Research* 144(1), 235-247.
4. Ben-Akiva, M., Bergman, M., Daly, A.J., Ramaswamy, R., 1984. Modeling inter-urban route choice behaviour, *Proceedings of the 9th International Symposium on Transportation and Traffic Theory*. VNU Science Press Utrecht, The Netherlands, pp. 299-330.
5. Ben-Akiva, M., Bierlaire, M., 1999. Discrete choice methods and their applications to short term travel decisions. *Handbook of transportation science* 23, 5-33.
6. Ben-Akiva, M.E., Lerman, S.R., 1985. *Discrete choice analysis: theory and application to travel demand*. MIT press.
7. Bhat, C.R., 2003. Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences. *Transportation Research Part B: Methodological*, 37 (9), 837-855.
8. Bierlaire, M., Frejinger, E., 2005. Route choice models with subpath components, *Swiss Transportation Research Conference*.
9. Biswas, M., Pinjari, A.R. and Dubey, S.K., 2019. Travel Time Variability and Route Choice: An Integrated Modelling Framework. In *11th International Conference on Communication Systems & Networks (COMSNETS)* (pp. 737-742). IEEE.
10. Bliemer, M., Bovy, P., 2008. Impact of route choice set on route choice probabilities. *Transportation Research Record: Journal of the Transportation Research Board* (2076), 10-19.
11. Bovy, P.H., 2009. On modelling route choice sets in transportation networks: a synthesis. *Transport reviews* 29 (1), 43-68.
12. Bovy, P.H., Fiorenzo-Catalano, S., 2007. Stochastic route choice set generation: behavioral and probabilistic foundations. *Transportmetrica* 3 (3), 173-189.
13. Broach, J., Gliebe, J., Dill, J., 2010. Calibrated labeling method for generating bicyclist route choice sets incorporating unbiased attribute variation. *Transportation Research Record: Journal of the Transportation Research Board* (2197), 89-97.
14. Cascetta, E., Nuzzolo, A., Russo, F., Vitetta, A., 1996. A modified logit route choice model overcoming path overlapping problems. Specification and some calibration results for interurban networks, *Internaional symposium on transportation and traffic theory*, pp. 697-711.

15. de la Barra, T., Perez, B., Anez, J., 1993. Multidimensional path search and assignment, *PTRC Summer Annual Meeting, 21st, 1993, University of Manchester, United Kingdom*.
16. Dhakar, N., 2010. Route choice modeling using GPS data. Doctoral Dissertation. University of Florida.
17. Dhakar, N., Srinivasan, S., 2014. Route choice modeling using GPS-based travel surveys. *Transportation Research Record: Journal of the Transportation Research Board* (2413), 65-73.
18. Dijkstra, E.W., 1959. A note on two problems in connexion with graphs. *Numerische mathematik* 1 (1), 269-271.
19. Feng, T., Arentze, T., Timmermans, H., 2013. Capturing preference heterogeneity of truck drivers' route choice behavior with context effects using a latent class model. *EJTIR* 13 (4), 259-273.
20. Fiorenzo-Catalano, S., Van Nes, R., Bovy, P.H., 2004. Choice set generation for multi-modal travel analysis. *European journal of transport and infrastructure research EJTIR, 4 (2)*.
21. Frejinger, E., Bierlaire, M., Ben-Akiva, M., 2009. Sampling of alternatives for route choice modeling. *Transportation Research Part B: Methodological* 43 (10), 984-994.
22. Frejinger, E., Bierlaire, M., 2007. Capturing correlation with subnetworks in route choice models. *Transportation Research Part B: Methodological* 41 (3), 363-378.
23. Fosgerau, M., Frejinger, E. and Karlstrom, A., 2013. A link based network route choice model with unrestricted choice set. *Transportation Research Part B: Methodological*, 56, 70-80.
24. Halldórsdóttir, K., Rieser-Schussler, N., Axhausen, K.W., Nielsen, O.A., Prato, C.G., 2014. Efficiency of choice set generation methods for bicycle routes. *European journal of transport and infrastructure research* 14 (4), 332-348.
25. Hartigan, J.A., 1975. *Clustering algorithms*. Wiley New York.
26. 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.
27. Hoogendoorn-Lanser, S., 2005. *Modelling travel behaviour in multi-modal networks*. *Ph.D. Thesis, Delft Institute of Techonology.*
28. Hoogendoorn-Lanser, S., Van Nes, R., 2004. Multimodal choice set composition: Analysis of reported and generated choice sets. *Transportation Research Record: Journal of the*

Transportation Research Board (1898), 79-86.

29. Jan, O., Horowitz, A.J. and Peng, Z.R., 2000. Using global positioning system data to understand variations in path choice. Transportation Research Record, 1725(1), pp.37-44.
30. Kamali, M., Ermagun, A., Viswanathan, K., Pinjari, A.R., 2016. Deriving Truck Route Choice from Large GPS Data Streams. *Transportation Research Record: Journal of the Transportation Research Board* (2563), 62-70.
31. Knorring, J., He, R., Kornhauser, A., 2005. Analysis of route choice decisions by long-haul truck drivers. *Transportation Research Record: Journal of the Transportation Research Board* (1923), 46-60.
32. Kuppam, A., J. Lemp, D. Beagan, V. Livshits, L. Vallabhaneni, and S. Nippani., 2014. Development of a Tour-Based Truck Travel Demand Model Using Truck GPS Data. Presented at 93rd Annual Meeting of the Transportation Research Board, Washington, D.C.
33. Lefebvre, N., Balmer, M., 2007. Fast shortest path computation in time-dependent traffic networks. *7th Swiss transport research conference, Ascona*, September 2007.
34. Lu, W., Scott, D.M., and Dalumpines, R. 2018. Understanding bike share cyclist route choice using GPS data: Comparing dominant routes and shortest paths. *Journal of Transport Geography* 71, 172-181.
35. Luong, T. D., Tahlyan, D., & Pinjari, A. R. 2018. Comprehensive Exploratory Analysis of Truck Route Choice Diversity in Florida. Transportation Research Record. https://doi.org/10.1177/0361198118784175
36. Papinski, D., Scott, D. M., & Doherty, S. T. 2009. Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. Transportation research part F: traffic psychology and behaviour, 12(4), 347-358.
37. Prato, C., Bekhor, S., 2006. Applying branch-and-bound technique to route choice set generation. *Transportation Research Record: Journal of the Transportation Research Board* (1985), 19-28.
38. 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.
39. Ramming, M.S., 2001. Network knowledge and route choice. *Ph. D. Thesis, Massachusetts Institute of Technology*.

40. Rieser-Schüssler, N., Balmer, M., Axhausen, K.W., 2013. Route choice sets for very high-resolution data. *Transportmetrica A: Transport Science* 9 (9), 825-845.
41. Rieser-Schüssler, N., Axhausen, K.W., 2009. Accounting for route overlap in urban and suburban route choice decisions derived from GPS observations, *12th International Conference on Travel Behaviour Research*, Jaipur.
42. Shams, K., Jin, X., Fitzgerald, R., Asgari, H., & Hossan, M. S. 2017. Value of Reliability for Road Freight Transportation: Evidence from a Stated Preference Survey in Florida. Transportation Research Record, 2610(1), 35-43.
43. Tahlyan, D., Luong, T.D., Pinjari, A.R., Ozkul, S., 2017. *Development and Analysis of Truck Route Choice Data for the Tampa Bay Region using GPS Data*. Report BDK25-730-3. Florida Department of Transportation.
44. Thakur, A., Pinjari, A.R., Zanjani, A.B., Short, J., Mysore, V., Tabatabaee, S.F., 2015. Development of Algorithms to Convert Large Streams of Truck GPS Data into Truck Trips. *Transportation Research Record: Journal of the Transportation Research Board* (2529), 66-73.
45. Ton, D., Duives, D., Cats, O. and Hoogendoorn, S., 2018. Evaluating a data-driven approach for choice set identification using GPS bicycle route choice data from Amsterdam. *Travel Behaviour and Society*, 13, pp.105-117.
46. Torrey, I., Ford, W., Murray, D., 2014. An analysis of the operational costs of trucking: 2014 Update, *American Transportation Research Institute.*
47. Zimmermann, M., Mai, T. and Frejinger, E., 2017. Bike route choice modeling using GPS data without choice sets of paths. *Transportation Research Part C: Emerging Technologies*, 75, 183-196.

Table 1: OD Pair-level Errors by Minimum No. of Observed Trips and Spatial Aggregation

Spatial Aggregation Level	Minimum Number of Trips	No. of OD Pairs	No. of Trips	No. of Observed Unique Routes		No. of Generated Unique Routes		False Negative Error		Weighted False Negative Error		False Positive Error	
				Mean	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*	*Mean*	*S.D.*
Link level	20	615	29,003	2.6	2.3	9.2	4.4	0.34	0.34	0.17	0.32	0.81	0.19
	30	335	22,327	2.8	2.4	8.9	4.5	0.38	0.35	0.19	0.35	0.81	0.19
	50	145	15,315	3.0	2.9	8.3	4.4	0.43	0.35	0.19	0.36	0.81	0.19
	100	48	8,995	3.4	2.8	7.2	4.5	0.53	0.33	0.26	0.41	0.79	0.2
XY cluster	20	1071	51,556	4.0	3.3	17.7	10.7	0.39	0.31	0.19	0.29	0.87	0.10
	30	615	40,654	4.6	3.6	18.3	11.2	0.44	0.29	0.18	0.28	0.87	0.10
	50	282	28,266	5.0	4.2	18.9	12.7	0.45	0.30	0.17	0.27	0.86	0.10
	100	80	15,008	6.2	5.4	19.9	14.3	0.55	0.24	0.19	0.29	0.86	0.09
Spatial cluster	20	966	58,774	5.5	4.3	26.0	20.1	0.41	0.29	0.18	0.25	0.87	0.09
	30	574	49,491	6.4	4.9	26.7	20.3	0.45	0.29	0.18	0.25	0.86	0.09
	50	294	39,001	7.4	5.7	28.0	19.8	0.49	0.27	0.18	0.26	0.86	0.10
	100	111	26,417	9.4	7.4	29.6	22.1	0.52	0.24	0.17	0.25	0.84	0.11
TAZ level (max. 2 km^2)	20	373	16,851	6.0	4.1	32.2	22.1	0.38	0.27	0.15	0.21	0.89	0.07
	30	205	12,989	6.8	4.5	32.6	22.6	0.43	0.26	0.14	0.19	0.88	0.07
	50	84	8,211	7.6	5.2	33.0	28.5	0.47	0.23	0.11	0.15	0.88	0.07
	100	28	4,336	8.3	6.2	33.4	28.4	0.54	0.21	0.11	0.18	0.88	0.08
TAZ level (max. 5 km^2)	20	723	40,229	6.8	4.7	36.9	28.4	0.38	0.26	0.17	0.22	0.88	0.07
	30	423	33,181	7.8	5.1	38.8	29.6	0.41	0.26	0.16	0.20	0.88	0.07
	50	196	24,602	8.9	5.8	39.2	27.1	0.44	0.23	0.14	0.19	0.87	0.07
	100	74	16,307	11.0	6.5	43.3	34.0	0.48	0.21	0.15	0.19	0.86	0.08
TAZ level (max. 10 km^2)	20	1152	70,494	7.7	5.8	41.4	33.2	0.38	0.25	0.18	0.23	0.88	0.08
	30	697	59,726	9.0	6.6	44.1	36.5	0.41	0.25	0.18	0.24	0.87	0.09
	50	336	46,047	10.7	7.8	47.6	38.0	0.44	0.24	0.17	0.23	0.87	0.09
	100	132	31,986	13.1	9.6	51.1	42.5	0.47	0.22	0.16	0.22	0.85	0.11

S.D. = standard deviation

Table 2: Comparison of Errors at Various Limits on Maximum Number of Routes to Generate in OD Pairs with at least 50 Trips at TAZ-Level (Max. Area = 2 Km^2) and Link-Level Aggregation

Limit on No. of Unique Routes at Link-level	Measure	TAZ Level (max. 2 km^2)				Link Level			
		No. of Generated Unique Routes	*False Negative Error*	*Weighted False Negative Error*	*False Positive Error*	*No. of Generated Unique Routes*	*False Negative Error*	*Weighted False Negative Error*	*False Positive Error*
5	Mean	21.10	0.49	0.11	0.83	4.50	0.45	0.20	0.75
	S.D.	10.23	0.23	0.16	0.09	0.97	0.35	0.37	0.20
10	Mean	27.90	0.47	0.11	0.86	7.04	0.43	0.19	0.80
	S.D.	16.75	0.23	0.15	0.07	2.91	0.35	0.36	0.19
15	Mean	32.16	0.47	0.11	0.88	8.28	0.43	0.19	0.81
	S.D.	22.11	0.23	0.15	0.07	4.44	0.35	0.36	0.19
20	Mean	36.19	0.46	0.11	0.88	8.59	0.42	0.19	0.81
	S.D.	25.19	0.23	0.15	0.07	4.98	0.35	0.36	0.19
No limit	Mean	37.56	0.46	0.11	0.89	8.68	0.42	0.19	0.81
	S.D.	26.69	0.23	0.15	0.07	5.24	0.35	0.36	0.19

S.D. = standard deviation

Table 3: Route Choice Model Estimated with TAZ Level (max. area = 2 km^2) Choice sets Aggregated from up to 5 BFS-LE Alternatives at Link Level

Variable Description	Error Components Logit with Random Parameter on Travel Time Variable	
	Parameter Estimate	***t-stat***
Travel cost ($)	-0.1261	-6.513
Travel time (min)	Mean = -0.097 Std. Dev = 0.6034	-3.003 30.635
Proportion of tolled portion of a route	-17.4014	-25.905
No. of turns per minute	-0.3996	-4.989
No. of ramps per minute	-0.2453	-2.489
Proportion of interstate portion of a route[φ]	36.3844	36.552
Proportion of major arterial portion of a route	22.3101	22.372
Proportion of minor arterial portion of a route	12.5747	15.432
Proportion of collector portion of a route	6.2076	8.089
Natural log of path size	-2.8777	-40.71
σ_{I-4}	2.3289	17.512
σ_{I-75}	2.2604	13.956
σ_{Polk}	1.397	9.986
σ_{US-19}	2.9823	2.72
No. of cases	6,453	
Log-likelihood at convergence	-9,681.31	
Log-likelihood for equal shares model	-19,327.52	
Rho-square	0.4991	
Adjusted rho-square	0.4983	

[φ] Roads were classified into one of 5 categories: interstate, major arterial, minor arterial, collector, and local road.

Table 4: Model Fit Measures for Route Choice Models Estimated Using Different Choice Sets

Model Specification	Model Fit Measures	Choice Set at Link Level with up to 5 BFS-LE Alternatives	Choice Set at Link Level with up to 15 BFS-LE Alternatives	Choice Set at TAZ Level (max. area = 2 km2) Aggregated from up to 5 BFS-LE Alternatives at Link Level	Choice Set at TAZ Level (max. area = 2 km2) Aggregated from up to 15 BFS-LE Alternatives at Link Level
Path Size Logit	$\mathcal{LL}_C$	-5,332.06	-6,915.12	-10,775.18	-11,970.52
	$\mathcal{LL}_{ES}$	-10,590.42	-15,951.96	-19,327.52	-21,674.72
	$\overline{\rho^2}$	0.496	0.566	0.442	0.447
	AIC	10,682.12	13,848.24	21,566.36	23,961.04
	BIC	10,672.89	13,839.01	21,559.13	23,949.81
Error Components Logit	$\mathcal{LL}_C$	-4,789.81	-6,303.43	-10,067.51	-11,331.78
	$\mathcal{LL}_{ES}$	-10,590.42	-15,951.96	-19,327.52	-21,674.72
	$\overline{\rho^2}$	0.546	0.604	0.478	0.477
	AIC	9,607.62	12,634.86	20,163.02	22,691.56
	BIC	9,588.39	12,615.60	20,143.79	22,672.33
Error Components Logit with Random Parameter on Travel Cost	$\mathcal{LL}_C$	-4,727.12	-6,129.55	-9,994.44	-11,229.98
	$\mathcal{LL}_{ES}$	-10,590.42	-15,951.96	-19,327.52	-21,674.72
	$\overline{\rho^2}$	0.552	0.615	0.482	0.481
	AIC	9,482.24	12,287.10	20,018.88	22,487.96
	BIC	9,463.01	12,267.87	19,997.65	22,468.73
Error Components Logit with Random Parameter on Travel Time	$\mathcal{LL}_C$	-4,609.86	--	-9,681.31	-10,810.01
	$\mathcal{LL}_{ES}$	-10,590.42	--	-19,327.52	-21,674.72
	$\overline{\rho^2}$	0.564	--	0.498	0.501
	AIC	9,245.72	--	19,392.62	21,648.02
	BIC	9,228.49	--	19,371.39	21,628.79

-- Model did not converge.

Table 5: Prediction performance of rotue choice models with various choice sets and model specifications

Model Specification	Measure of expected overlap	Choice Set at Link Level with up to 5 BFS-LE Alternatives	Choice Set at Link Level with up to 15 BFS-LE Alternatives	Choice Set at TAZ Level (max. area = 2 km2) Aggregated from up to 5 BFS-LE Alternatives at Link Level	Choice Set at TAZ Level (max. area = 2 km2) Aggregated from up to 15 BFS-LE Alternatives at Link Level
Path Size Logit	Mean (std. dev)	0.9290 (0.0741)	0.9340 (0.0737)	0.9192 (0.0722)	0.9190 (0.0743)
Error Components Logit	Mean (std. dev)	0.9130 (0.0734)	0.8203 (0.1878)	0.8018 (0.2752)	0.7913 (0.3530)
Error Components Logit with Random Parameter on Travel Cost Variable	Mean (std. dev)	0.9135 (0.0735)	0.8204 (0.1880)	0.8017 (0.2751)	0.7914 (0.3487)
Error Components Logit with Random Parameter on Travel Time Variable	Mean (std. dev)	0.9136 (0.0746)	--	0.8016 (0.2752)	0.7914 (0.3527)

-- Model did not converge. Hence it was not used for evaluation.

Table 6: Comparison of Route Characteristics of Observed and Generated Routes in OD Pairs with at least 50 Trips at TAZ Level (Max. Area = 2 Km^2) Aggregation

Route Characteristics	Relevant Routes Captured in Generated Choice Sets (i.e., Observed and Generated)		Irrelevant/Extraneous Routes (i.e., Generated but not Observed)	
	Mean	***Std. Dev.***	***Mean***	***Std. Dev.***
Length (mi)	43.350	22.360	45.050	22.640
Proportion of ramps	0.037	0.039	0.049	0.034
Proportion of tolled roads	0.000	0.062	0.028	0.063
Proportion of interstate highways and major arterials	0.784	0.284	0.667	0.255
Proportion of minor arterials	0.137	0.222	0.173	0.190
Proportion of collectors	0.061	0.105	0.131	0.101
Proportion of local roads	0.018	0.040	0.0290	0.047
No. of links	214.90	123.920	253.200	119.100
No. of links per mile	5.750	3.070	6.460	2.820
No. of intersections	89.770	77.010	119.300	72.510
No. of intersections per mile	2.580	2.070	3.220	1.960
No. of right turns	1.950	1.520	4.750	2.260
No. of left turns	1.920	1.290	4.850	2.480
Average path size	0.29*(0.09)#	0.19(0.06)	0.140	0.060

*Pathsize of observed relevant routes w.r.to observed routes. #Pathsize of generated relevant routes w.r.to generated routes.

APPENDIX

Comparison of OD Pair-level Evaluation Results to Trip-level Evaluation Results

The errors reported in Table 1 are OD pair-level errors, as opposed to trip-level errors typically reported in the literature. The trip-level error computed out of all 82,738 trips used in this study is 0.25 (i.e., observed routes for 25% of the trips were not present in the generated choice sets). When we examined only those trips belonging to OD pairs with a minimum of 20 trips at various spatial aggregations, the corresponding trip-level errors ranged from 0.18 for all 16,851 trips between TAZs of up to 2 km^2 size to 0.28 for all 58,774 trips between spatial clusters. It is interesting to note that both the trip-level errors and OD pair-level average errors are smallest for the spatial aggregation of TAZs (of up to 2 km^2 size).

The trip-level false negative errors from various studies in the literature that use repeated shortest path based choice set generation methods are reviewed in Table A1 for different tolerance thresholds on the difference between observed and generated routes. Although it is difficult to compare errors reported in different studies due to differences in the travel modes, choice set generation algorithms, and specifics of implementation, one can observe from the reported errors of the current study and another study by Hess et al. (2015) that the use of BFS-LE approach to generate route choice sets for truck travel seems to result in relatively small trip-level errors compared to that for other modes of travel. To examine this further, we analyzed (for all 82,738 trips used in Table 1) how different are the observed routes from their corresponding shortest time routes and shortest distance routes on the network. More than 80% of the observed routes had commonality factors above 0.9 with respect to their corresponding shortest time route. Further, only about 70% of the observed routes had commonality factors above 0.9 with respect to their corresponding shortest distance route. This shows that the chosen routes by trucks are not very different from the shortest time and shortest distance route in an OD pair and perhaps this is the reason behind the performance of BFS-LE approach, which is based on repeated shortest time search. This finding is different from the work by Jan et al. (2000) who report the observed routes to be significantly different from shortest time paths in case of car travel. This suggests that the performance of BFS-LE algorithm in generating route choice sets for cars and other modes of travel might not be as good as it is for freight trucks. Another reason the current study had a small error rate (compared to other studies) is because we generated up to 15 *unique* route alternatives

(at the link-level) that were different from each other by at least 5%. Most other studies consider generated routes as different from each other even if they are different from each other by a small link and generate up to a maximum of 15 or 20 such routes (which are not very different from each other). This limits the diversity of generated routes and, therefore, limits the capture of diverse observed routes.

Evaluation of Generated Choice Sets at Different Thresholds of Overlap between Observed and Generated Choice Sets

In all the analysis presented in the paper, the generated unique choice sets were compared to the observed unique choice sets using a threshold value of 0.95 for the commonality factor. Table A2 provides weighted false negative and false positive errors computed for OD pairs with a minimum of 50 trips at the spatial aggregation of TAZ-level (of up to 2 km^2) for different thresholds values of commonality factors—0.95, 0.90. 0.85, and 0.80. It can be observed that the weighted false negative error values decreased substantially as the threshold value decreased — an average false negative error of 0.11 at 0.95 threshold value to an average false negative error of 0.04 at 0.90 threshold value. Admittedly, threshold values of 0.90 or more are a bit too high for trips of mid-rage to long distance. However, the results suggest that most uncaptured observed routes (with a 0.95 threshold value) are not substantially different from the generated routes, highlighting the performance of the BFS-LE algorithm implemented in this paper.

Table A1: False Negative Errors for Various Choice Set Generation Algorithms

Algorithm	Study	Mode	Max. Number of Alternatives	Important Features of Used Generation Algorithm	False Negative Error (%)		
					Tolerance (%)		
					0	*10*	*20*
Breadth-first-search link elimination	Present study	Truck	15**	Use of free-flow travel time as cost function to generate routes that are at least 5 percent different from each other.	25 (at 5% tolerance)		
	Rieser-Schüssler et al. (2013)	Car	20*	Use of free-flow travel time as cost function	37	N.T.	N.T.
			100*		27	N.T.	N.T.
	Hess et al. (2015)	Truck	15*	Use of generalized cost function that includes penalties that reflect other sources of inconvenience occurring on minor roads	26	N.T.	N.T.
	Halldórsdóttir et al. (2014)	Bicycle	20*	Use of generalized cost function taking into account road types, cycle lanes, and land use	34	28	22
	Ton et al. (2018)	Bicycle	20*	Use of distance as travel cost	99	98	97
	Dhakar and Srinivasan (2014)	Car	20**	Use of commonly factor to generate routes that are at least 5% different from each other	N.T.	51	N.T.
Link elimination	Bekhor et al. (2006)	Car	N.R.	Elimination of links on shortest path (in sequence) to generate new routes	40	37	29
	Prato and Bekhor (2007)	Car	10*	Elimination from shortest path of links that takes driver farther from destination and closer to origin or compels driver to turn from high hierarchical road to low hierarchical road	42	42	30
`Labeling	Bekhor et al. (2006)	Car	3*	Generation of routes to minimize distance, free-flow time. and time	61	56	48
			16*	Use of 16 different labels to generate various routes	28	24	15
	Prato and Bekhor (2007)	Car	4*	Generation of routes to minimize distance, free-flow time, travel time, and delay	60	60	60
	Broach et al. (2010)	Bicycle	9*	Use of 11 different labels to generate various routes but still making sure that no generated route deviate from shortest path by more than 100%	80	75	65
	Ton et al. (2018)	Bicycle	N.R.	Use of various labels to generate routes	99	98	96
Calibrated labeling	Broach et al. (2010)	Bicycle	20*	Generation of routes using multiple labels and cost function parameters, calibrated using observed distribution of shortest path deviation	78	71	58
Link penalty	Bekhor et al. (2006)	Car	40*	Shortest route generation after gradual increase of impedance of all links on shortest path	43	33	20
			15*		44	34	22
	Prato and Bekhor (2007)	Car	15*	Iterative shortest route generation after increasing impedance of shortest path by factor of 1.05	46	46	38
Simulation (low variance)	Prato and Bekhor (2007)	Car	N.R	Generation of shortest path by drawing link impedances from truncated normal distribution with mean travel to travel time, variance equal to 20% of mean, left truncation limit equal to free-flow travel time, right truncation limit equal to time for speed of 10km/h	51	51	46
Simulation (high variance)	Prato and Bekhor (2007)	Car	N.R	Generation of shortest path by drawing link impedances from truncated normal distribution with mean travel to travel time, variance equal to 100% of mean, left truncation limit equal to free-flow travel time, right truncation limit equal to time for speed of 10km/h	39	38	29
Doubly stochastic generation function	Fiorenzo-Catalano et al. (2004)	Multi-modal	1600*	Repeated shortest path generation by considering stochasticity in travelers' perception of network attributes and preferences for different trip components	22	N.T.	N.T.

N.R: Maximum number of generated alternatives not reported in study.
N.T: Particular tolerance level not tested in study.
** Generated route alternatives were elemental alternatives (i.e. two route alternatives considered separate alternatives even if they differ from each other by one link.)*
*** Generated alternatives were unique alternatives (i.e. two route alternatives considered separate alternatives if they differ from each other by a certain minimum non-overlap.*

Table A2: Comparison of Errors at Various Overlapping Thresholds in OD Pairs with at Least 50 Trips at TAZ Level (Max. Area = 2 km^2) Aggregation

Overlapping Threshold Value	Weighted False Negative Error		False Positive Error	
	Mean	S.D.	Mean	S.D.
0.95	0.11	0.15	0.88	0.07
0.9	0.04	0.08	0.79	0.14
0.85	0.02	0.07	0.76	0.17
0.8	0.01	0.03	0.74	0.20