Comprehensive Exploratory Analysis of Truck Route Choice Diversity in Florida

Trang D. Luong
Department of Civil and Environmental Engineering
University of South Florida
4202 East Fowler Ave, Tampa, FL 33620, USA
Tel: +1-813-408-3927; Email: trangluong@mail.usf.edu

Divyakant Tahlyan
Department of Civil and Environmental Engineering
University of South Florida
4202 East Fowler Ave, Tampa, FL 33620, USA
Tel: +1-813-508-2598; Email: dtahlyan@mail.usf.edu

Abdul R. Pinjari (Corresponding Author)
Department of Civil Engineering
Center for Infrastructure, Sustainable Transportation and Urban Planning (C*i*STUP)
Indian Institute of Science (IISc)
Bangalore 560012, India
Tel: +91-80-2293 2043; Email: abdul@iisc.ac.in

Word count: 6,493 words text + 4 tables/figures x 250 words (each) = 7,493 words

TRB Paper number: 18-05667

Submission Date: March 1, 2018

ABSTRACT
This study presents a comprehensive exploratory analysis of truck route choice diversity in the state of Florida, for both long-haul and short-haul truck travel segments. We employ six metrics to measure three different dimensions of diversity in truck route choice between any given origin-destination (OD) pair. These dimensions are: (a) number of distinct routes used to travel between the OD pair, (b) the extent of overlap (or lack thereof) among the routes, and (c) the evenness (or the dominance) in the usage of different unique routes. The diversity metrics were applied to a large database of 73,000 truck routes derived from 200 million GPS records. Descriptive analysis and statistical modeling of the diversity metrics offered insights on the determinants of various dimensions of truck route choice diversity between any OD pair. The results are useful for improving choice set generation algorithms for truck route choice modeling and in truck route policies and investments.

Keywords: truck route choice, route diversity, route overlap, route dominance, truck-GPS data

INTRODUCTION

An essential step toward enhancing highway freight mobility is to improve our understanding of freight-truck route choice behavior. Specifically, analysis of the routes that trucks use to travel between different origins and destinations can support the design of truck routing policies aimed at mitigating congestion, improving travel time reliability, and facilitating truck movement during network disruptions. However, research on truck route choice has been limited due to inadequacy of data on truck movements. The recent availability of global positioning systems (GPS) data has encouraged research on truck route choice modeling *(1,2,3)* and highway freight performance measures *(4,5)*. Yet, little attention has been paid to exploring the diversity or variability of truck route choice between travel origins and destinations.

An improved understanding of truck route diversity has significance in both freight modeling and planning. For modeling applications, understanding truck route choice diversity can help determine the number and structure (i.e., extent of overlap) of route alternatives to be used in route choice models and traffic assignment procedures *(6)*. For planning applications, analyzing the diversity of truck route choices observed in the field can help inform truck routing decisions during regular and emergency situations. For example, identifying origin-destination (OD) pairs with high travel demand but low diversity in the routes used (e.g., a single route used) can help identify critical segments of the network and inform routine infrastructure maintenance scheduling as well as re-routing efforts during emergency recovery. Also, one can apply route diversity measures to evaluate the redundancy of (or lack thereof) truck routes in existing transportation infrastructure to justify long-term investments on truck corridors to increase network redundancy.

This research presents a comprehensive exploratory analysis of truck route choice diversity in the state of Florida for both long-haul and short-haul travel segments. Specifically, the paper addresses two broad questions: (a) How to measure the degrees of diversity in the routes trucks use to travel between an OD pair? (b) What factors influence the diversity of truck route choice between an OD pair? To this end, six metrics are utilized to measure the following three different dimensions of diversity in route choice between a given OD pair: 1) number of different routes used between the OD pair, 2) extent of overlap (or lack thereof) among the routes, and 3) evenness (or the dominance) of the use of different unique routes between that OD pair. These metrics are applied to quantify truck route choice diversity using a database of about 73,000 routes derived from more than 200 million truck-GPS records. Next, statistical models are estimated to explore the influence of various determinants on the three dimensions of route choice diversity between hundreds of OD pairs. The models provide insights into the influence of the characteristics of truck travel demand, OD location, and network structure on the diversity of truck route choice between an OD pair. Potentially, these insights can help travel modelers in improving choice set generation algorithms for modeling truck route choice and help planners in devising truck routing policies.

The next section reviews past studies on variability in route choice behavior. The following section summaries the truck-GPS data used for this study. The metrics used to quantify diversity in truck route choice are then elaborated. The next section presents the statistical modelling methodologies used in this study. Subsequently, empirical results are presented and discussed. The last section concludes the paper.

LITERATURE

Only a handful of studies in the literature have analyzed route diversity. Jan et al. *(6)* were among the first to analyze variability in observed routes, using GPS data from a household survey in Lexington, Kentucky. They concluded that most chosen routes differed from the shortest-time path and that the extent of variability across trips made by different households was greater than that across trips made within the same household. In another study, Papinski et al. *(7)* estimated that up to 20% of travelers deviated from their planned route. They found that the observed routes are longer than both shortest-time and shortest-distance routes because travelers consider many route characteristics in addition to time and distance. Additionally, Spissu et al. *(8)* examined the intra-personal and inter-personal variability in route choice using GPS data of 679 routes chosen by a sample of 12 university students. Their results suggested higher levels of intra-individual variability (than inter-individual variability) for discretionary trips and higher levels of inter-individual variability for work or study trips. Finally, Zhu and Levinson *(9)* utilized a 13-day GPS data of 143 individuals in the Minneapolis–St. Paul Twin Cities to analyze the diversity of commute trips to investigate their hypothesis that people choose a portfolio of routes over time.

In all these studies, the datasets were limited to only tens or hundreds of observed routes. The metrics used by these studies for analyzing route variability cannot be used to measure the amount of overlap among routes or dominance of usage of routes. Furthermore, none of the above studies are focused on truck route choice. The current paper fills such gaps in research through a comprehensive analysis of truck route diversity using six different metrics of diversity applied on a large database of truck routes. In addition, this is perhaps the first empirical study that examines the influence of truck travel demand, OD location, and network structure on diversity of truck route choice. Besides, a comparison of route diversity patterns between long-haul and short-haul travel segments is presented.

DATA

The truck-GPS data used in this study was obtained from the American Transportation Research Institute (ATRI). The database of long-haul truck trips (trips longer than 50 miles) comprised 145 million GPS records collected from nearly 50,000 freight trucks that traveled in the state of Florida from March to June 2010. The data used to derive short-haul trips (trips shorter than 50 miles) comprised about 96 million GPS records corresponding to 110,000 freight trucks and spanned six counties of the Tampa Bay region in Florida. Temporally, the short-haul data corresponded to the first 15 days in October 2015, December 2015, April 2016, and June 2016.

The raw GPS data were first converted into a database of truck-trips using algorithms developed by *(10)* and later refined by *(11)*. To derive the chosen route for each trip, raw GPS records corresponding to each trip were map-matched to high-resolution NAVTEQ roadway networks provided by the Florida Department of Transportation (FDOT), using the procedure developed by *(12)*. The 2010 NAVTEQ network used to derive long-haul routes comprised 1.5 million links and 5.8 million nodes whereas the 2015 NAVTEQ network used for short-haul routes comprised 1.8 million links and 6.9 million nodes. After map-matching and validation process, a database of about 78,000 long-haul trips and 225,000 short-haul trips was retained.

To analyze route choice and the diversity therein, it is useful to aggregate trip end locations to larger spatial units. This spatial aggregation allows analysts to observe sufficient number of trips to get an uncensored view of the various routes trucks choose between an OD pair. Hence, all trip end locations in this study were aggregated to the traffic analysis zones (TAZs) defined in Florida Statewide Travel Demand Model. TAZ OD pairs with at least 50 long-haul trips and at least 30 short-haul trips were selected, since OD pairs with fewer trips might not offer a complete picture of truck route diversity. The final long-haul dataset used in this analysis comprises 277 TAZ OD pairs with a total of 30,263 trips longer than 50 miles. The short-haul dataset comprises 527 TAZ OD pairs with a total of 42,884 trips of 5 to 50 miles length. Trips shorter than 5 miles were not considered in this analysis because such trips would typically not have many route choice options.

DIVERSITY METRICS

To measure diversity in truck route choice between any given OD pair, we employ the following six metrics: 1) Number of unique routes, 2) Average commonality factor, 3) Average path size, 4) Non-overlapping index, 5) Standardized variance of route usage, and 6) Standardized Shannon entropy of route usage. The first metric measures the number of unique routes traveled by trucks between an OD pair. The next three metrics measure the extent of overlap (or lack thereof) among the unique routes identified between that OD pair. The last two metrics measure the evenness (or, otherwise, dominance) in the use of unique routes between the OD pair. These three dimensions together provide a complete picture of the diversity in truck route choice between any OD pair. All six metrics are defined next.

Number of Unique Routes

Many routes chosen between an OD pair overlap substantially with each other and are different by only a few links. To determine a set of distinct or unique routes based on the amount of overlap, we used the commonality factor (C_{ij}) between routes i and j, proposed by *(13)* as follows:

$$C_{ij} = l_{ij}/\sqrt{L_i L_j} \qquad (1)$$

where

L_i and L_j = the lengths of routes i and j, respectively; and
l_{ij} = the length of the shared portion between the two routes.

The two routes are defined as unique if the commonality factor between the two routes is less than 0.95. To determine the number of unique routes observed between an OD pair, all routes between that OD pair are arranged in an ascending order of route length. The shortest route is the first unique route. The commonality factor of each subsequent route is computed with respect to all preceding unique routes to determine whether it is a unique route (if all C_{ij} values for that route are less than 0.95). The result of this process is a set of unique routes between an OD pair, where the C_{ij} value between any two unique routes is less than 0.95. The size of this unique route set represents the number of unique routes used between that OD pair.

Average Commonality Factor

Since the number of unique routes metric does not quantify the extent of overlap (or lack thereof) among the unique routes observed between an OD pair, we derive the average commonality factor metric to measure the degree of overlap between all pairs of unique routes between the OD pair, as shown below:

$$\overline{\overline{CF}} = \frac{\sum_{1}^{\binom{K}{2}} C_{ij}}{K} \qquad (2)$$

where

C_{ij} = the commonality factor computed for two unique routes i and j between an OD pair;
K = the number of unique routes between the OD pair; and
$\binom{K}{2}$ = the number of pairs of unique routes.

Ranging between 0 and 1, an average commonality factor value closer to 0 represents low overlap between the unique routes. Therefore, for a given number of unique routes between an OD pair, a lower value of $\overline{\overline{CF}}$ suggests a greater extent of diversity (i.e., less overlap) among the routes.

Average Path Size

Proposed by *(14),* path size (PS) is a metric commonly used in the route choice modeling literature to measure the degree of overlap of two routes between an OD pair. The PS for each unique route i is defined as follows:

$$PS_i = \sum_{a \epsilon \Gamma_i} \left(\frac{l_a}{L_i}\right) \frac{1}{\sum_{j \epsilon unique\ routes} \delta_{aj}} \tag{3}$$

where

Γ_i is the set of all links composing unique route i;
l_a is the length of link a;
L_i is the length of route i; and
δ_{aj} is equal to 1 if route j belonging to the set of K unique routes uses link a, 0 otherwise.

The maximum possible value of PS is 1 and the minimum value tends to zero. A route with no overlap with any other routes has a PS value of 1. Average PS in an OD pair is the mean value of PS across all unique routes between that OD pair. For a given number of unique routes between an OD pair, the closer to 1 is the average PS value, the greater is the diversity (i.e., lower overlap) among routes.

Non-overlapping Index

Complementary to the above two metrics, the degree of non-overlap among the unique routes between an OD pair is quantified using the non-overlapping index. This index is measured as the ratio between the total length of links (on unique routes) that were used only once to the total length of all links (on unique routes) that were used at least once. This index ranges between 0 and 1, where a value closer to 1 represents low overlap among unique routes.

Standardized Variance of Route Usage

Another dimension of diversity is based on the evenness in the usage of different unique routes between an OD pair. The most even usage is when all observed trips are equally distributed among the observed unique routes between that OD pair. A complementary concept is the degree of dominance, when most trips are observed to have taken only one or a few unique routes.

To measure the degree of evenness of route usage, the distribution of *N* trips among *K* different unique routes between a given OD pair may be characterized as a multinomial distribution, with each trip being allocated to any one of the K different unique routes. If the random variable X_k $(k = 1,2,3, \dots, K)$ indicates the number of trips choosing route k and p_k is the proportion of trips allocated to route k, vector $X = (X_1, X_2, \dots, X_K)$ follows a multinomial distribution with parameters N and p, where $p = (p_{1,}p_2, \dots, p_K)$. The variance of such multinomial-distributed random variables may be written as $Var(X_k) = N * p_k * (1 - p_k)$.

The variance of route usage between an OD pair is defined as the sum of variances of usage frequency for each route, as: $N * \sum_1^K p_k * (1 - p_k)$. This metric is influenced by three factors that vary across OD pairs: 1) number of observed trips (more trips, higher the variance); 2) number of unique routes (more routes, higher the variance); and 3) evenness of the distribution of the observed trips among various unique routes. For a given OD pair with *N* observed trips and *K* unique routes, the maximum possible value of variance of route usage, when all trips are evenly distributed among all unique routes, is determined as follows:

$$Max\ variance\ of\ route\ usage = N * K * (1/K) * (1 - 1/K) = N * (1 - 1/K) \qquad (4)$$

where

N = the number of observed trips between an OD pair; and
K = the number of unique routes between the OD pair.

To measure solely the nature of trip distribution without being influenced by the number of observed trips *(N)* or unique routes *(K)*, this metric may be standardized as the ratio of the variance of usage to the maximum possible variance as shown in Equation 5:

$$Standardized\ variance\ of\ route\ usage = \frac{\sum_1^K p_k * (1 - p_k)}{(1 - \frac{1}{K})} \qquad (5)$$

where

K = the number of unique routes between an OD pair; and
p_K = proportion of trips taking the k^{th} unique route.

The closer this metric is to its maximum possible value 1, the more evenly distributed the observed trips are among various unique routes. For example, if 100 trips use two unique routes between an OD pair, the standardized variance of usage for that OD pair would be 1 if 50 trips take the first route and the other 50 trips take the second route.

Standardized Shannon Entropy of Route Usage

Shannon entropy *(15)* is a metric typically used to measure the evenness of distribution of different entities among a given number of categories. Proposed in the field of information science, the concept of entropy has been applied widely in different fields, such as geodiversity and biodiversity, and used in transportation to measure land use diversity. We apply the same concept for route choice diversity to measure the extent of distribution of all trips among different unique routes. The Shannon entropy of usage of *K* unique routes between an OD pair is defined as $\sum_1^K p_k \ln(p_k)$. The maximum value of the Shannon entropy of usage is $K * (1/K) * \ln(1/K) = \ln(1/K)$ when all trips are equally distributed among the identified unique routes between an OD pair. To eliminate the effect of number of unique routes, the standardized Shannon entropy (SE) of route usage is computed as below:

$$Standardized\ SE\ of\ route\ usage = \frac{\sum_1^K p_k \ln(p_k)}{\ln(1/K)} \qquad (6)$$

The maximum possible value Standardized SE is 1, when all trips are evenly distributed among all unique routes.

An Illustration

Figure 1 illustrates the application of the above diversity metrics for two different OD pairs. The first OD pair is selected from the long-haul data, with 8 unique routes that are 62 to 84 miles long. Note that many of the 8 unique routes overlap considerably with each other. Such overlap is measured by the average commonality factor and average path size. 53 out of the 66 trips observed between this OD pair use the first unique route, indicating the dominance of the first unique route. The second OD pair is from short-haul data with 32 observed trips that are more evenly distributed among the different routes than those between the first OD pair. Such differences in dominance (or evenness) of route usage are measured by the two metrics–standardized variance of route usage and the standardized Shannon entropy of route usage (Figure 1).

MODELING METHODOLOGY

Here, we explain the statistical model structures used to analyze the determinants of the following three of the six metrics developed for each OD pair: 1) number of unique routes, 2) average path size of unique routes, and 3) standardized Shannon entropy of route usage.

Count Data Models for Number of Observed Unique Routes

Negative binomial (NB) regression is an appropriate choice to model count data given by the number of observed unique routes in this research. Typically, Poisson regression is preferred if the mean of the count process is equal to the variance. If there is a significant difference between the mean and the variance of the count process, the data are said to be over-dispersed and NB regression is preferred. Our empirical data supported the use of NB regression over Poisson regression, because of over-dispersion in the data.

In NB regression, the probability $P(y_i)$ of an OD pair i having y_i number of unique routes is defined as follows:

$$P(y_i) = \frac{\Gamma(1/\alpha + y_i)}{\Gamma(1/\alpha) y_i!} \left(\frac{1/\alpha}{(1/\alpha)+\lambda_i}\right)^{1/\alpha} \left(\frac{y_i}{(1/\alpha)+\lambda_i}\right)^{y_i} \quad (7)$$

where

$\Gamma(\cdot)$ = the gamma function:

$\lambda_i = exp(\beta X_i + \varepsilon_i)$;

X_i = a vector of explanatory variables;

β = a vector of parameters to be estimated; and

$exp\ (\varepsilon_i)$ is a Gamma-distributed disturbance term with unit mean and variance given by the dispersion parameter α.

The model parameters can be estimated using maximum likelihood estimation technique. Depending on the count process being modeled, the regression can be right, left, or two-side truncated. To model the number of unique routes between an OD pair, our count data models were left-truncated at 1, because any OD pair in the data has at least one unique route.

Fractional Response Models for Average Path Size and Standardized Shannon Entropy of Usage

It is worth noting that all diversity metrics proposed in this study, except the number of unique routes, range between 0 and 1. The fractional response model structure proposed by *(16)* may be used to model such quantities whose values lie between 0 and 1. Although proportion data may be modeled by logit transformation of the dependent variable [$i.e.,\ \ln(y_i/(1-y_i) = \beta X_i$] followed by ordinary least squares regression, this transformation cannot be used when the dependent variable might take values of zero or one. This issue can be resolved with the fractional response model *(19)* whose expected value of the dependent variable is: $E(y_i|x_i) = G(x_i\beta)$. Two possible functional forms for $G(z)$ are: (a) logistic function, $G(z) = \exp(z)/(1+\exp(z))$ and (b) cumulative density function of a standard normal distribution. According to this model, the quasi likelihood of an OD pair with an observed value y_i is given as below:

$$\mathcal{L}_i(\beta) = y_i * \log[G(x_i\beta)] + (1-y_i) * \log[1-G(x_i\beta)] \quad (8)$$

where

$G(x_i\beta)$ = a known function with $0 < G(z) < 1\ \forall z \in \mathbb{R}$.

The parameters are estimated using maximization of the quasi log-likelihood function.

DESCRIPTIVE ANALYSIS

The first set of rows in Table 1 summarizes the mean and standard deviation (SD) values of all diversity metrics calculated for our long-haul and short-haul datasets. There are 19 long-haul OD pairs and 22 short-haul OD pairs that have one single observed unique route. Except for the number of unique routes metric, all other diversity metrics reported were computed after excluding such OD pairs with only one unique route (Table 1).

It is worth noting that the short-haul routes exhibit greater diversity than long-haul routes with higher average values for observed unique routes, non-overlapping index, standardized variance of usage and standardized Shannon entropy of usage. The standard deviations are also higher for short-haul data, suggesting a greater incidence of higher values. In other words, short-haul routes are move diverse than long-haul routes from the standpoint of lower overlap as well as lower dominance in their usage.

Potential Determinants of Diversity

To explore the correlates of diversity in truck route choice, we extracted a variety of factors describing observed travel demand, OD locations, and network structure between all OD pairs. These variables are presented in the second set of rows in Table 1 and discussed next.

Trip Characteristics

The first category of variables includes the number of trips observed for each OD pair and the number of trucks taking those trips (a measure of truck travel demand), spatial separation (straight-line distance or direct distance) between the OD locations, and travel conditions measured between the OD pair (particularly on the most used route). In terms of travel conditions of the most used route, we measure travel time variability and level of route circuity (defined as the ratio of route length to the direct OD distance).

OD Location Characteristics
Characteristics of origin and destination TAZs include land-use descriptors (employment densities, TAZ size, urban/rural classification) and spatial dispersion of freight activity centers (defined as the average distance of all trip ends' centroid to each trip end location).

Network Structure
To explore the impact of network structure on the diversity of observed routes, we hypothesized two different areas of influence between OD pairs. In the first hypothesis, the diversity of route choice between an OD pair is provided by the entire road network inside an elliptical area of influence connecting that OD pair, referred to as the long ellipse. The long ellipse's major axis is assumed to be the same distance and orientation of the straight line connecting the centroids of origin and destination TAZs. Its minor axis is set to be one-third of the major axis length. In the second hypothesis, the diversity of route choice between an OD pair is differentially impacted by two different areas of influence. The first category of influence areas are circular areas around the origin and destination TAZ centroids, referred to as circular buffers. The buffer radii explored are 1, 2, and 5 miles for direct distances of [5-10), [10-20), and more than 20 miles, respectively. The second area of influence is elliptically shaped, referred to as the short ellipse, with the major axis as the difference of straight line distance and radius of the circular buffers on each end.

Within these hypothesized areas of influence for each OD pair, we computed densities of various road types (major arterials, minor arterials, collectors, and local roads) to characterize the network structure between the OD pair. In addition, other facilities along the roadway, such as traffic signals, intersections, interchanges, and truck rest stops, were counted within long and short ellipses and circular buffers.

STATISTICAL MODELING RESULTS
Statistical models were estimated separately for long-haul and short-haul datasets to analyze the determinants of diversity metrics—number of unique routes, average path size, and standardized Shannon entropy. Such empirical model results are discussed here.

NB Regression Model for Number of Unique Routes
Table 2 presents the NB regression estimation results for the number of unique routes, separately for long-haul and short-haul travel segments. Both model results indicate that OD pairs with a higher number of observed trucks are likely to have more unique routes. This is an expected result because more trucks traveling between an OD pair may lead to greater diversity in route choice due to heterogeneity in preferences of truck drivers, operators, and the businesses they serve. Similarly, OD pairs with more observed trips have more unique routes, in both travel segments (specifically, when there are more than 150 trips in the short-haul segment). More trips suggest a greater demand for travel and may lead to greater diversity in route choices as well. The next variable in the long-haul model, indicating high travel time variability (when the difference between 95^{th} and 5^{th} percentile travel time on the most used route is greater than 15 minutes), suggests more unique routes since the high variability (or low reliability) in travel conditions causes truckers to prefer alternative routes. Specific to the long-haul model, deviation of the most used route from the straight-line OD distance (measured as the ratio of the most used route length to straight-line distance) has a positive influence on the number of observed unique routes. When the most used route is more circuitous, more available routes in the network may exist and increase truckers' considerations for more alternate routes. Interestingly, neither travel time variability nor route circuity had a significant influence in the short-haul segment.

In the context of OD location characteristics, OD pairs with larger OD TAZs are likely to have lower number of unique routes in both travel segments, perhaps because those TAZs are typically in areas with sparse network, population, and employment density, and therefore fewer network options to travel. For the same reason, both OD locations being in an urban zone is associated with a higher number of unique routes in the short-haul model. In the context of direct distance, OD pairs that are more than 200 miles apart tend to have less unique routes. This finding is consistent with the limited options roadway network of Florida offers for long-haul travel.

We now discuss the influence of OD location variables that are significant only in the short-haul model. The number of unique routes increases as spatial separation increases from small (<10 miles) to moderate (10-20 miles), and then decreases in the highest length segment. This may be because the road network in the Tampa Bay region does not offer many route options both for local (<10 miles) as well as cross-county (>40 miles) travel. Employment densities at the OD TAZs are positively correlated with the number of unique routes; perhaps because a greater employment density is a surrogate for the heterogeneity of businesses served by freight trucks, which leads to a greater diversity in route choice. Similarly, average distance between the TAZ-centroid of all trip ends to each trip end (a measure of spatial dispersion of the freight activity generators in the OD TAZs) is positively associated with the number of unique routes. The average distance from TAZ-centroids to the nearest major arterial is a surrogate for how far trucks need to travel to reach a major arterial, which is negatively correlated with the number of unique routes.

In the context of network structure, long-haul OD pairs with higher ratio of toll roads to major arterials captured in the long ellipse are likely to have more unique routes; which can be explained by truck operators' decision to look for alternative routes to avoid tolls *(2)*. This variable is, however, insignificant in the short-haul model mostly because our study region for the short-haul segment does not have many toll roads. In the long-haul model, OD locations with higher density of major and minor arterials in circular buffers around trip ends are likely associated with lower number of unique routes while OD locations with higher density of collectors are likely associated with more unique routes. This may be because the availability of more major and minor arterials around origins or destinations reduces the need to search for alternative routes. On the other hand, OD pairs with higher density of major and minor arterials and collectors in the short ellipse are likely to have more unique routes; probably because of an increased number of route options. For similar reasons, OD pairs with a greater proportion of minor arterials and collectors (with respect to major and minor arterials and collectors) in the short ellipse are likely to have a greater number of observed unique routes in the long-haul model. In the short-haul model, whereas the density of the minor arterials and collectors in the long ellipse has a positive influence on the number of observed unique routes, the influence of the proportion of major arterials (with respect to major, minor arterials and collectors) is negative. All these results highlight subtle but notable differences in the influence of network structure on the diversity of truck route choice between long-haul and short-haul travel segments (Table 2).

Fractional Response Models for Average Path Size and Standardized Shannon Entropy of Usage

Table 3 presents the fractional response model estimation results for average path size (first set of rows) and standardized Shannon entropy of usage (second set of rows), estimated for OD pairs with at least two observed unique routes (Table 3).

The average path size models have results expected for both long-haul and short-haul segments. OD pairs with more observed unique routes are likely to have lower average path size (i.e., greater overlap). OD pairs with a higher proportion of trips on the most used route are likely to have higher average path size (i.e., lower overlap). The presence of a dominant route may imply the presence of other longer routes that do not overlap much and are less preferable. A greater spatial separation of OD pairs is associated with a smaller value of path size (i.e., greater overlap) of the different unique routes; perhaps because an increase in spatial separation may reduce the number of travel routes offered by the network.

As expected for both long-haul and short-haul travel, when modeling standardized Shannon entropy, OD pairs with more observed unique routes are likely to have a higher Shannon entropy (i.e., more even distribution of trips among unique routes). OD pairs with more observed trips are likely to have a more even usage of the routes. OD pairs with a higher average path size value (or lower overlap) among unique routes demonstrate a more uneven usage of different routes. Such OD pairs with less overlapping routes are likely to have one or few dominant routes that are largely preferred over other routes. In the short-haul model, OD pairs with a greater average distance from the centroid of the trip end TAZs to all trip ends (i.e., greater spatial dispersion of freight activity generators) are likely to be associated with a more even distribution of trips among different unique routes. This suggests the influence of heterogeneity or spatial dispersion in trip ends on the heterogeneity of preferences for truck routes.

SUMMARY AND CONCLUSIONS

This study presents a comprehensive exploratory analysis of truck route choice diversity in Florida, for both long-haul and short-haul travel segments. To measure the diversity in truck routes between any OD pairs, we develop the following six metrics: 1) Number of unique routes, 2) Average commonality factor, 3) Average path size, 4) Non-overlapping index, 5) Standardized variance of route usage, and 6) Standardized Shannon entropy of route usage. The first metric measures the number of distinct truck routes between an OD pair. The next three metrics measure the extent of overlap (or lack thereof) among the routes. The last two metrics measure the evenness (or, otherwise, the dominance) of the usage of the routes between the OD pair. These three dimensions together provide a complete picture of the diversity in truck route choice between an OD pair. The diversity metrics were utilized to describe truck route choice diversity in Florida from a database of about 73,000 truck trips derived from more than 200 million GPS records. We produced a rich database of diversity metrics for 277 TAZ OD pairs for long-haul travel (trips longer than 50 miles) in the State of Florida and 527 TAZ OD pairs for short-haul travel (trips between 5 to 50 miles) in the Tampa Bay region. In addition, we complied an extensive set of variables characterizing the truck travel, OD location, and network structure between these OD pairs that could potentially influence the extent of route choice diversity. Negative binomial regression models were estimated to explore the influence of these variables on the number of unique routes traveled between an OD pair. Further, fractional response models were estimated to explore the determinants of average path size (overlap among routes) and standardized Shannon entropy (evenness) of route usage.

The analysis suggests that short-haul truck travel exhibits greater diversity in route choice than long-haul travel, in terms of number of unique routes observed, the extent of non-overlap between unique routes, as well the evenness of usage of different unique routes. Within the long-haul segment, OD pairs that are more than 200 miles apart exhibit lower diversity than those that are closer. Among the short-haul OD pairs, local (<10 miles) and cross-county (>40 miles) travel exhibit lower diversity than medium distance travel. OD pairs in urban zones are associated with a greater diversity in route choice, because urban areas offer more network options for route choice. OD pairs with a greater number of trips and/or trucks observed (i.e., greater demand for travel) are associated with a higher number of unique routes. OD pairs with more variability in travel conditions (e.g.: travel time) or routes that deviate more from a straight-line have more diverse traveled routes. In addition, network structure between OD pairs has a considerable influence on the diversity of truck route choices. Another important finding is that the determinants and their extent of influence differs between short-haul and long-haul travel segments. For example, OD TAZ land-use (employment density and diversity of freight activity locations) has a significant influence on route choice diversity only in the short-haul segment. Furthermore, network structure variables have differential impacts on route diversity between the two segments. Finally, OD pairs with a higher number of observed unique routes have greater overlap (i.e., lower average path size) and lower dominance of route usage, whereas OD pairs with less overlapping routes exhibit greater dominance of usage.

The findings from this study can be applied to improve algorithms for generating choice sets for truck route choice modeling. Specifically, route choice set generation algorithms can be customized based on truck travel demand, OD location, and network structure characteristics found to be influential in this analysis. Furthermore, the count models of number of unique routes can be used to determine the appropriate number of choice alternatives to be generated for route choice modeling and traffic assignment purposes.

An enhanced understanding of route diversity can also help improve truck routing policies and determine routing decisions during emergency situations. For example, OD pairs with only one unique route but high travel demand tend to be vulnerable in the event of roadway disruptions. Thus, transportation planners can focus on such OD pairs to build a proactive plan for emergency detours as well as better maintenance of critical road segments. Furthermore, such OD pairs can be priority candidates for future roadway expansion to create additional route options.

ACKNOWLEDGEMENTS

This study was funded by FDOT and the Center for Advanced Infrastructure and Transportation (CAIT), a USDOT's University Transportation Center. The opinions, findings, and conclusions expressed here are not necessarily those of FDOT or USDOT. Three anonymous reviewers provided constructive comments on an earlier manuscript.

REFERENCES

1. Knorring J.H., He R., Kornhauser A.L. Analysis of Route Choice Decisions by Long-haul Truck Drivers. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1923, 2005, pp. 46-60.
2. Wang Z, Goodchild A. GPS Data Analysis of the Impact of Tolling on Truck Speed and Routing: Case Study in Seattle, Washington. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2411, 2014, pp. 112-119.
3. Hess S., Quddus M., Rieser-Schüssler N., Daly A. Developing Advanced Route Choice Models for Heavy Goods Vehicles Using GPS Data. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 77, 2015, pp. 29-44.
4. Wang Z., Goodchild A., McCormack E. Measuring Truck Travel Time Reliability Using Truck Probe GPS Data. *Journal of Intelligent Transportation Systems,* Vol. 20, No. 2, 2016, pp. 103-112.
5. Liao C-F. Generating Reliable Freight Performance Measures with Truck GPS Data: Case Study in Twin Cities Metropolitan Area, Minnesota. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2410, 2014, pp. 21-30.
6. Jan O., Horowitz A., Peng Z-R. Using Global Positioning System Data to Understand

Variations in Path Choice. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1725, 2000, pp. 37-44.

7. Papinski, D., Scott, D., Doherty, S. 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*, Vol. 12, 2009, pp. 347-358.
8. Spissu, E., Meloni I., Snjust, B. Behavioral Analysis of Choice of Daily Route with Data from Global Positioning System. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2230, 2011, pp. 96-103.
9. Zhu, S., Levinson., D. A Portfolio Theory of Route Choice. *Transportation Research Part C: Emerging Technologies*, Vol. 35, 2013, pp. 232-243.
10. Thakur, A., Zanjani, A. B., Pinjari, A. R., Short, J., Mysore, V., Tabatabaee, S. F. Development of Algorithms to Convert Large Streams of Truck GPS Data into Truck Trips. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2529, 2015, pp. 66-73.
11. Pinjari A.R., Kamali M., Luong T.D., Ozkul S. *GPS Data for Truck-Route Choice Analysis of Port Everglades Petroleum Commodity Flow.* Report BDK25-977-17. Florida Department of Transportation, 2015.
12. Kamali M., Ermagun A., Viswanathan K., Pinjari A.R. Deriving Truck Route Choice from Large GPS Data Streams. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2563, 2016, pp. 62-70.
13. Cascetta, E., Nuzzolo, A., Russo, F., Vitetta, A. A Modified Logit Route Choice Model Overcoming Path Overlapping Problems. Specification and Some Calibration Results for Interurban Networks. *International Symposium on Transportation and Traffic Theory*, 1996, pp. 697-711.
14. Ben-Akiva M., Bierlaire M. Discrete Choice Methods and Their Applications To Short Term Travel Decisions. *Handbook of Transportation Science*, Vol. 23, 1999, pp. 5-33.
15. Shannon C.E. A Mathematical Theory Of Communication. *Bell System Technical Journal.* Vol. 27, 1948, pp. 379–423.
16. Papke L.E., Wooldridge J. Econometric Methods For Fractional Response Variables with an Application to 401 (K) Plan Participation Rates. *Journal of Applied Econometrics*, Vol. 11, No. 6, 1996, pp. 619-632.

LIST OF TABLES

LIST OF FIGURES

TABLE 1 Descriptive Statistics of All Data Used for Route Diversity Analysis

No.	Diversity Metrics		Long-haul		Short-haul	
			Average	SD	Average	SD
1	Number of unique routes		8.61	6.54	9.03	6.51
2	Average commonality factor		0.69	0.17	0.68	0.18
3	Average path size		0.28	0.12	0.29	0.14
4	Non-overlapping index		0.26	0.13	0.32	0.15
5	Standardized variance of usage		0.62	0.26	0.65	0.25
6	Standardized Shannon entropy of usage		0.57	0.21	0.61	0.21
No.	**Potential Determinants of Diversity**		**Long-haul**		**Short-haul**	
			Average	**SD**	**Average**	**SD**
1	No. of trips observed for an OD pair		109.3	99.7	81.1	123.7
2	No. of trucks observed for an OD pair		19.7	16.6	68.3	113.5
3	OD airway distance (miles)		109.8	69.6	21.9	11.0
4	Travel time of trips taking the most used route (minutes)	SD	7.5	3.6	4.8	18.4
5		Average	146.1	73.9	31.4	13.0
6		95th percentile	158.5	77.4	35.9	14.3
7		5th percentile	136.9	71.7	27.8	12.1
8	Ratio of the most used route length to airway OD distance		1.2	0.1	1.2	0.4
9	Employment density of OD TAZs (1000 jobs/sq. mile)	All types	7.0	5.5	5.7	4.0
10		Industrial	1.5	1.1	0.9	0.7
11		Service	3.4	3.1	3.0	2.8
12		Commercial	2.1	1.9	1.8	1.3
13	Average area of OD TAZs (miles^2)		2.2	2.8	2.3	3.4
14	Indicator if both OD TAZs are urban		0.8	0.3	0.9	0.3
15	Average distance from centroid of all trip ends to each trip end (miles)		0.4	0.5	0.8	1.0
16	Average distance from TAZ centroid to major arterials (miles)		6.0	3.5	4.7	3.2
17	Length of major arterials (miles)*	Long ellipse	331.8	411.8	27.1	27.9
18		Short ellipse	274.2	386.1	15.7	17.5
19		Ending buffers	29.0	23.6	8.8	10.0
20	Length of minor arterials (miles)*	Long ellipse	621.4	855.4	42.6	49.4
21		Short ellipse	502.6	800.1	18.2	25.9
22		Ending buffers	62.8	60.2	17.8	24.0
23	Length of collectors (miles)	Long ellipse	1276.6	1561.8	110.5	114.3
24		Short ellipse	1031.3	1446.2	48.7	62.2
25		Ending buffers	137.9	96.0	48.1	56.7
26	Length of local roads (miles)	Long ellipse	10155.1	14471.0	673.1	639.2
27		Short ellipse	9867.2	11496.5	311.3	349.2
28		Ending buffers	1212.1	768.9	259.8	279.2
29	Length of toll roads (miles)	Long ellipse	81.4	117.6	5.8	8.2
30	No. of rest stops	Long ellipse	9.2	12.1	1.0	1.4
31		Short ellipse	8.1	11.6	0.6	1.0
32		Ending buffers	0.3	0.9	0.4	0.8
33	No. of interchanges	Long ellipse	84.4	124.6	9.3	12.1
34		Short ellipse	60.5	107.6	4.0	6.7
35		Ending buffers	18.1	24.3	4.2	7.1
36	No. of traffic signals	Long ellipse	728.6	1043.8	59.2	86.3
37		Short ellipse	532.9	901.4	25.0	47.4
38		Ending buffers	136.2	168.4	29.1	53.0

Note: ∗the lengths of major and minor arterials include both toll roads and toll-free roads

TABLE 2 Estimation Results of Truncated Negative Binomial Regression of Number of Unique Routes for Long-haul and Short-haul Datasets

Variable Description	Long-haul Data		Short-haul Data	
	Coefficient	t-stat	Coefficient	t-stat
Trip Characteristics				
Logarithm of number of truck IDs	0.293	5.53	0.358	10.93
Logarithm of number of trips	0.379	4.97	--	--
Trip indicator (1 if more than 150 trips, 0 otherwise)	--	--	0.159	1.80
Travel time variability on the most used route indicator (1 if the difference of 95^{th} and 5^{th} percentile of travel time greater than 15 minutes, 0 otherwise)	0.172	1.86	--	--
Ratio of length of the most used route to direct OD distance (miles/mile)	1.486	3.51	--	--
OD Characteristics				
Average area of OD TAZs ($mile^2$)	-0.040	-1.85	-0.036	-3.85
Indicator if both OD TAZs are urban zone	--	--	0.277	3.83
Indicator if direct OD distance between 50-200 miles	0.381	3.59	N/A	N/A
Indicator if direct OD distance between 10-20 miles	N/A	N/A	0.128	2.84
Indicator if direct OD distance more than 40 miles	N/A	N/A	-0.671	-6.52
Industrial employment density (1000 jobs/$mile^2$)	--	--	0.215	6.41
Commercial employment density (1000 jobs/$mile^2$)	--	--	0.059	2.95
Average distance from centroid of all trip ends to each trip end (miles)	--	--	0.283	12.54
Average distance from TAZ centroids to nearest major or minor arterials (miles)	--	--	0.042	4.87
Network Structure				
Ratio of toll roads to major arterials in long ellipse (miles/mile)	1.283	3.09	--	--
Density of major and minor arterials in 5-mile buffers around both endings (miles/$mile^2$)	-0.583	-2.87	--	--
Density of collectors in 5-mile buffers around both endings (miles/$mile^2$)	0.340	2.29	--	--
Density of major, minor arterials and collectors in short ellipse (miles/$mile^2$)	0.231	2.12	--	--
Density of minor arterials and collectors in long ellipse (miles/$mile^2$)	--	--	0.108	2.70
Proportion of major arterials to total length of major, minor arterials and collectors in long ellipse (miles/mile)	--	--	-0.534	-2.41
Proportion of minor arterials and collectors to total length of major, minor arterials and collectors in short ellipse (miles/mile)	1.375	2.24	--	--
Constant	-3.997	-5.06	-0.254	-1.41
Dispersion parameter	0.235	5.78	0.075	5.80
Number of observations (OD pairs)	277		527	
Log likelihood at convergence	-782.90		-1372.45	
Log likelihood for constant-only model	-842.92		-1617.90	
Adjusted ρ^2 with respect to constant-only model	0.056		0.141	

Note: For variables that have a significant influence in one model but not in the other model, "--" appears in place of parameter estimate and t-stat for that variable in the latter model. "N/A" used when variable is not applicable to a specific model.

TABLE 3 Estimation Results of Fractional Response Models for Average Path Size and Standardized Shannon Entropy of Usage

Variables in Average Path Size Model	Long-haul Data		Short-haul Data	
	Coefficient	t-stat	Coefficient	t-stat
Number of unique routes	-0.068	-9.66	-0.056	-9.74
Proportion of trips on the most used route	0.534	3.72	0.705	6.25
Direct OD distance (miles)	--	--	-0.012	-5.71
Direct OD distance indicator (1 if more than 200 miles, 0 otherwise)	-0.175	-1.95	--	--
Constant	-0.665	-4.73	-0.539	-4.70
Number of observations (OD pairs)	258		505	
Log pseudo likelihood at convergence	-101.08		-202.11	
Log pseudo likelihood for constant-only model	-106.37		-211.61	
Rho-square with respect to constant-only model	0.050		0.045	

Variables in Standardized Shannon Entropy Model	Long-haul Data		Short-haul Data	
	Coefficient	t-stat	Coefficient	t-stat
Number of unique routes	0.059	5.76	0.052	6.37
Number of trips	-0.003	-4.83	-0.002	-3.74
Average path size	-1.345	-1.78	-1.667	-4.31
Average distance from centroid of all trip ends to each trip end (miles)	--	--	0.255	4.60
Constant	0.498	1.81	0.462	2.80
Number of observations (OD pairs)	258		505	
Log pseudo likelihood at convergence	-119.19		-224.05	
Log pseudo likelihood for constant-only model	-126.39		-241.98	
Rho-square with respect to constant-only model	0.057		0.074	

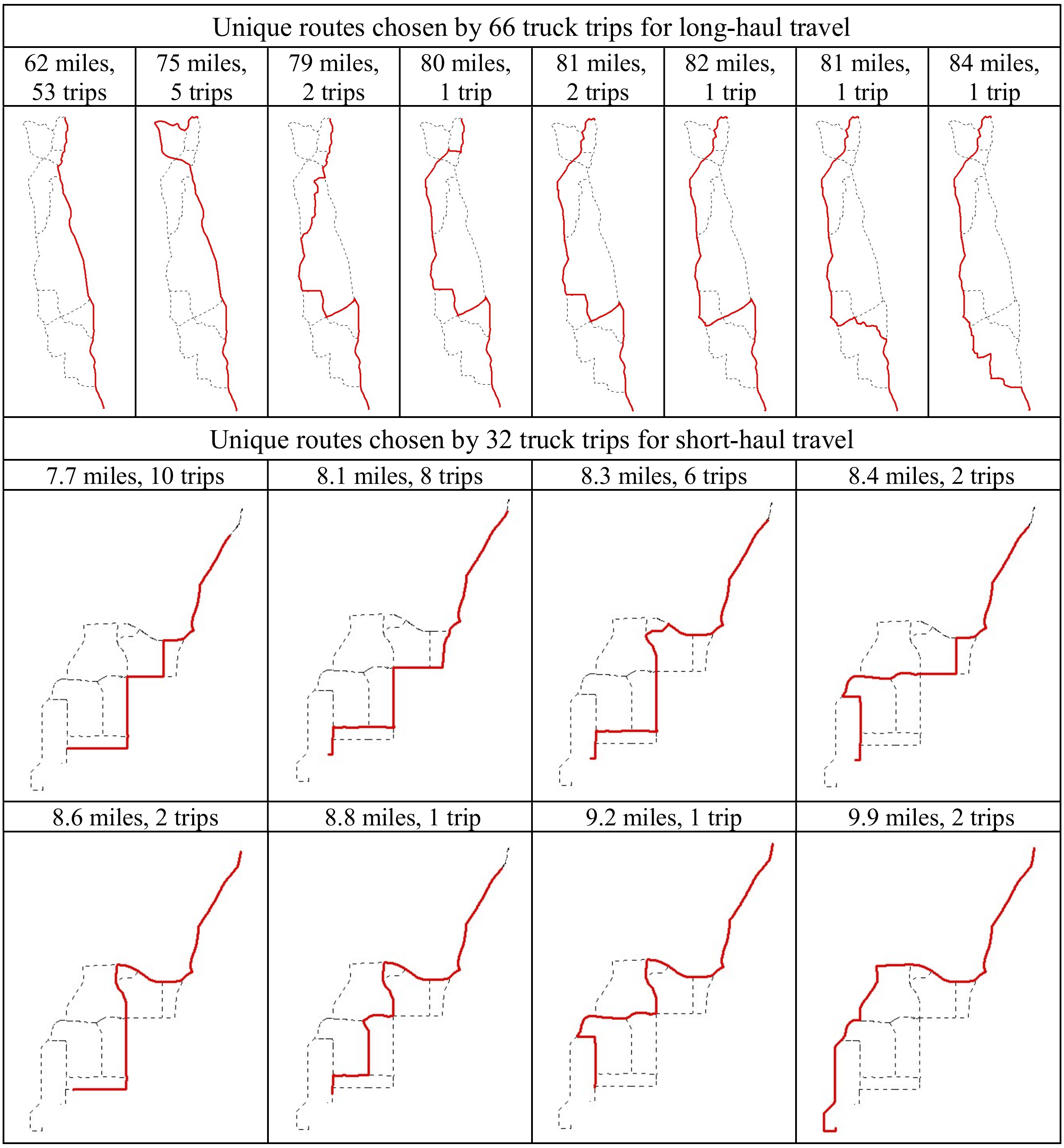

FIGURE 1 Examples of unique routes *(in bold)* for long-haul and short-haul OD pairs