Estimating Route Choice Characteristics of Truck Drivers from Sparse Automated Vehicle Identification Data through Data Fusion and Bi-Objective Optimization

Sharma, Salil; van Lint, Hans; Tavasszy, Lóránt; Snelder, Maaike

DOI
10.1177/03611981221095089

Publication date
2022

Document Version
Final published version

Published in
Transportation Research Record

TRR
JOURNAL OF THE TRANSPORTATION RESEARCH BOARD

Research Article

Transportation Research Record
2022, Vol. 2676(12) 280–292

DOI: 10.1177/03611981221095089
journals.sagepub.com/home/trr
SAGE

Estimating Route Choice Characteristics of Truck Drivers from Sparse Automated Vehicle Identification Data through Data Fusion and Bi-Objective Optimization

Salil Sharma[1], Hans van Lint[1], Lóránt Tavasszy[1], and Maaike Snelder[1,2]

Abstract

Optimizing route choices for truck drivers is a key element in achieving reliable road freight operations. For commercial reasons, it is often difficult to collect freight activity data through traditional surveys. Automated vehicle identification (AVI) data on fixed locations (e.g., Bluetooth or camera) are low-cost alternatives that may have the potential to estimate route choice models. However, in cases where these AVI sensors are sparsely located, the resulting data lack actual route choices (or labels), which limits their application estimating route choice models. This paper overcomes this limitation with a new two-step approach based on fusing AVI and loop-detector data. First, a sparse Bluetooth data set is fused with travel times estimated from densely spaced loop-detector data. Second, the combined data set is fed into a bi-objective optimization method which simultaneously infers the actual route choices of truck drivers between an origin–destination pair and estimates the parameters of a route choice (discrete choice-based) model. We apply this approach to investigate the route choice behavior of truck drivers operating to and from the port of Rotterdam in the Netherlands. The proposed model can distinguish between peak and off-peak periods and identify different segments of truck drivers based on a latent classes choice analysis. Our results indicate the potential of traffic and logistics interventions in improving the route choices of truck drivers during peak hours. Overall, this paper demonstrates that it might be possible to estimate route choice characteristics from readily available data that can be retrieved from traffic management agencies.

Keywords

data and data science, freight movement data, freight traffic, road freight vehicles (trucks), freight systems, driver, general

Road transport has been the main choice for inland freight transport within the European Union, accounting for 76.30% of the modal share in 2019. Especially in the Netherlands, where the port of Rotterdam generates most of the freight activity, the share of road freight was estimated at over 50% in 2019 (*1*). This reliance on road transport calls for robust and reliable traffic operations. On the one hand, freight transport contributes to congestion and, on the other, trucking companies in the Netherlands have suffered economic damage caused by road congestion. This economic damage is estimated to be €1.5 billion for 2019 and this cost has been increasing yearly (*2*). Therefore, a thorough investigation of on-trip route choices of truck drivers is fundamental to our understanding of how road freight moves. This, in turn, can support the development of advanced traffic and logistics interventions.

The estimation of route choice models requires data that are typically collected using either stated-preference (SP) or revealed-preference (RP) surveys. The pros and cons of SP and RP-based approaches are widely known. SP studies solicit choice behavior in hypothetical scenarios where the actual choices might be different than those

[1]Transport and Planning, Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, The Netherlands
[2]TNO, The Hague, The Netherlands

Corresponding Author:
Salil Sharma, S.Sharma-4@tudelft.nl

stated. RP studies rely on rich activity data sets and do not have these validity limitations. However, RP data cannot be collected under the same rigorously controlled circumstances as SP data (*3*). Nonetheless, for studes of the route choices of truck drivers there has been a recent shift from SP-based (*4–8*) to RP-based studies given the availability of trajectory (mostly GPS) data sets (*3*, *9–14*). Although GPS data are appealing because of their spatial (i.e., location) and temporal (i.e., timestamps) richness to investigate route choices, a few limitations are associated with collection and coverage. These data are often not publicly available and are expensive to buy from service providers. Furthermore, they might not capture a representative sample of the population over a limited period. In contrast to GPS data, there are other low-cost alternatives such as automated vehicle identification (AVI) data. These data are collected from fixed-location sensors (e.g., Bluetooth sensors or traffic cameras), which can be installed by road authorities on many different strategically chosen locations. These fixed-location sensors can alleviate the limitations of GPS data in two ways. First, these sensors can capture passing vehicles' movements to produce a large sample that is (more) representative of the population. Second, they can continuously record vehicles' movements over several periods of the day.

Fixed-location sensors have some advantages over mobile sensors (i.e., GPS). However, little research effort has been put into harnessing the potential of such data for route choice modeling (*15*). The key reason for that is that these sensors may not fully cover a road network sufficiently to make the underlying route choice observable (in the mathematical sense, e.g., Viti et al. [*16*]). The result typically is a sparse data set that comprises origin, destination, and experienced travel time for a given trip. This sparse data set is unlabeled in the sense that it lacks the actual route choices of drivers and, as such, cannot be used to estimate discrete-choice models. To deal with the sparsity issue, extra information about the estimated travel times of route alternatives is required to infer the most likely chosen route, that is, the missing label. One possibility is to use another independent data set (e.g., loop-detector data, floating car data) to derive this information. The estimation problem relies on the inference of the most likely chosen route, and this inference can be approached from the following two perspectives.

1. The most likely route chosen by a driver will maximize their perceived utility.
2. The most likely route chosen by a driver will minimize the deviation between experienced and estimated travel times.

Note that the deviation of travel times is computed from two independent data sets and might be associated with some uncertainty (*17*). A naïve approach that assigns missing labels based on the lowest deviation value might, therefore, produce erroneous estimates of model parameters. In contrast, Cao et al. (*15*) combine the aforementioned two perspectives into a single objective function, based on the so-called network-free model (*18*), to model route choices using camera (also sparse data) and GPS data. Although their approach is promising, it strongly depends on the quality of the available GPS data and how representative they are for the population. Moreover, their method incorporates the second perspective through a measurement equation to supply prior beliefs, which come from distributional assumptions, about a route present in the choice set.

Motivated by these issues, we propose an alternative approach that fuses a sparse Bluetooth data set with path travel times derived from densely spaced loop detectors. To estimate how long a trip would take on alternative routes, we use a trajectory-based travel time estimation approach (*19*). In this way, our approach does not depend on GPS data and their variability. In addition, the estimation problem is investigated in a bi-objective optimization setting that allows the capture of the interdependency between the conflicting perspectives: utility maximization and deviation minimization. Therefore, this approach can be used to simultaneously infer actual route choices (labels) and estimate the parameters of a route choice (discrete choice-based) model under minimal assumptions. As a result, this approach is applied to estimate route choice characteristics of truck drivers operating in the Netherlands.

Turning now to route choice phenomena among truck drivers, the existing literature has studied time-of-day impacts and the latent class segmentation in SP-based contexts (*4*, *5*, *7*) where full experimental control is exerted by researchers and the data may suffer from hypothetical bias. The study of these two effects is particularly important for road freight for two reasons. First, it can provide us with insights into the vulnerability of road freight operations, especially in peak hours. Second, latent class choice models, unlike mixed logit models (*14*), do not require knowledge of any mixing distribution, thus making them more useful for policy and decision makers in the logistic and traffic sectors. This indicates a need to study these effects using route choices made by truck drivers in real-world situations. This paper fills this research gap by using a Bluetooth data set. Please note that this paper does not use data collected from either SP or RP surveys.

This paper aims to estimate the route choice characteristics of truck drivers using a sparse AVI or Bluetooth data set that lacks actual route choices. This paper contributes to the existing literature by:

1. estimating the route choice characteristics of truck drivers from a sparse AVI data set, where

actual route choices are lacking, in combination with loop-detector data through bi-objective optimization; and
2. investigating time-of-day effects and latent segmentation within route choices of truck drivers from Bluetooth data that include their decisions in real-world situations.

This paper is structured as follows. The first section will describe an approach to building a database of truck drivers and route-specific attributes using a sparse Bluetooth data set and loop-detector data. The subsequent section is concerned with the methodology where the bi-objective optimization approach and latent class modeling approach are described. The paper then presents the modeling results and discusses key findings. Finally, in the conclusions, we outline future research directions. Note that route and path are used interchangeably in this paper.

Data

This section first describes an approach to building a Bluetooth data set for truck drivers that can be used for modeling their route choices. It then presents the attributes of route alternatives necessary to capture the route choice behavior of truck drivers.

Bluetooth Data Set for Trucks

Bluetooth stations record the time stamp and identity of a passing vehicle equipped with a Bluetooth sensor. The identity is captured in the form of a media access control (MAC) address. The travel time between two Bluetooth stations can be retrieved by comparing the timestamps. For this paper, the Bluetooth data are provided by the Bluetooth service from the port of Rotterdam. This is a query-based service that returns data in the JavaScript object notation (JSON) format. This service ensures privacy by masking the real MAC address. However, Bluetooth data, in general, do not provide information related to vehicle type. This paper uses a three-step approach to prepare a Bluetooth data set for trucks. First, we identify pairs of Bluetooth stations that can be used to identify the vehicle type. Second, we prepare a database of truck drivers by storing their hashed MAC IDs. Finally, these MAC IDs are used to identify a truck trip from Bluetooth data. These trips provide key information (such as origin–destination [O-D] pair, and trip duration) that is necessary to estimate their route choice characteristics.

Identification of Bluetooth Stations that can Cluster Travel Time. Clustering has been used in the past to infer vehicle type by analyzing travel times between two Bluetooth sensors (*12*, *20*). In our data, we have found two pairs of Bluetooth stations near the ring of Rotterdam (A15 and A4) where each pair comprised one main Bluetooth station and one ancillary Bluetooth station (Figure 1). These pairs can cluster travel time observations in both travel directions, resulting in four sections for our analysis.

Identification of Truck Drivers in the Bluetooth Data Set. The method for extracting truck-specific data from the Bluetooth data set is presented in Figure 2. For a given day and a pair of Bluetooth stations, congested periods from the data set are removed since vehicles are observed to behave similarly as shown by travel time plots. Outliers are then removed using a quartile-based method (*21*). Next, we apply the Gaussian mixture model-based technique (*22*) to cluster travel time observations into one of the two groups: faster and slower vehicles. Note that the slower vehicles group might contain some of the slower passenger cars. We, therefore, use the regulatory speed limit of trucks on motorways in the Netherlands, that is, 80 km/h as a filter to remove undesired passenger cars and label the rest of the vehicles as "likely to be trucks." We iterate over different days (October and November 2017) and one of the four sections that can cluster travel time observations. After this process, we label vehicles that are found in the "likely to be trucks" category more than 90% of the time and are detected at least three times by any pair of the Bluetooth stations as "trucks." This process results in a database of hashed MAC IDs that represent truck drivers.

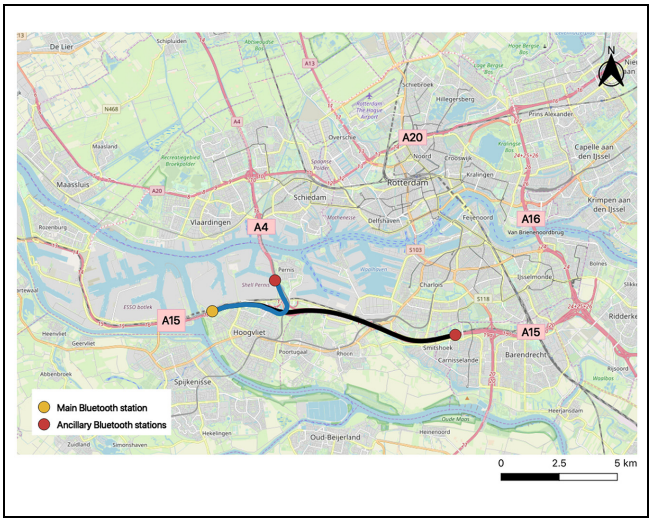

Figure 1. Two pairs of Bluetooth stations that can cluster travel times in both traffic directions.

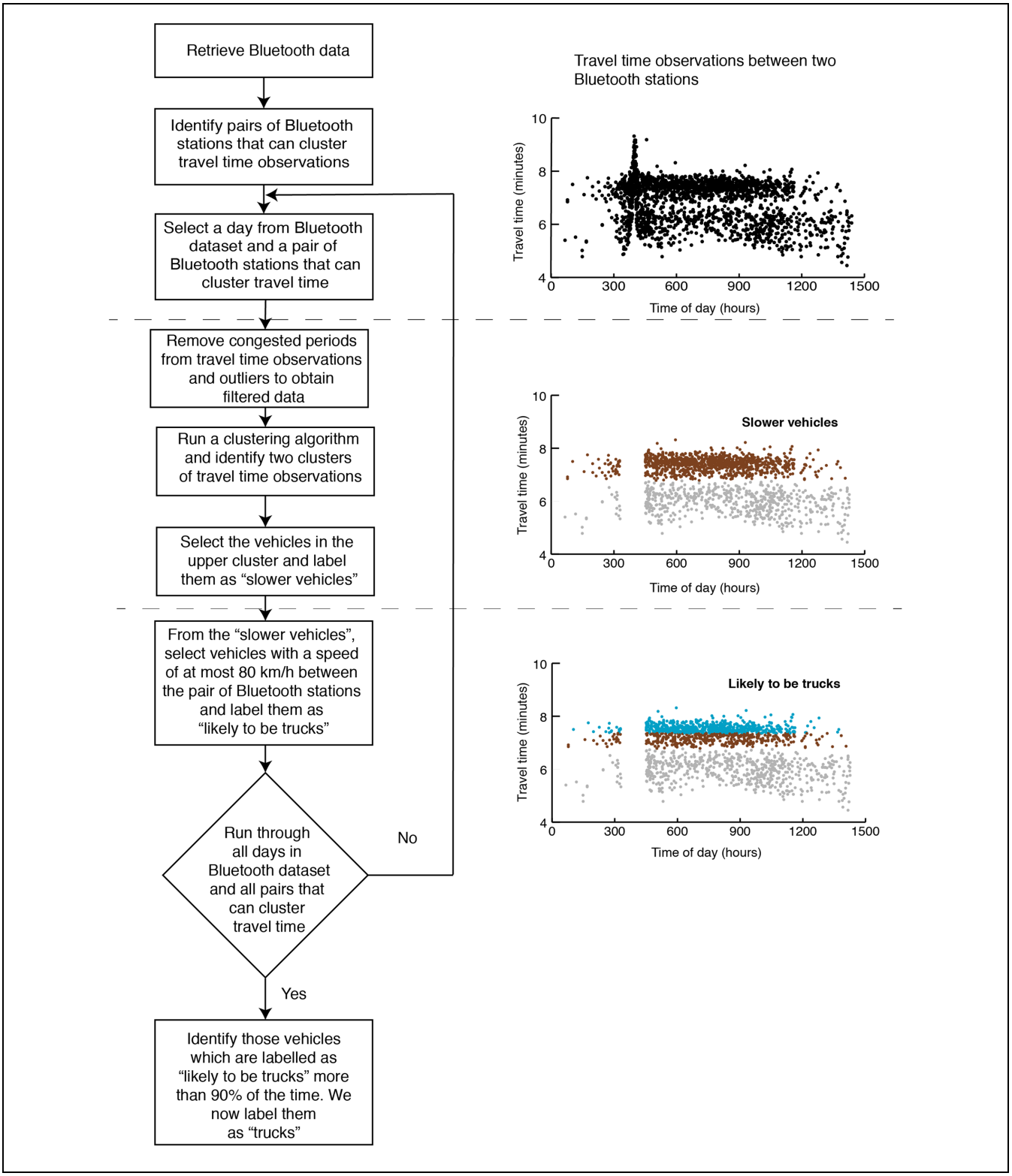

Figure 2. Extracting trucks-specific observations from Bluetooth data set.

Identification of Truck Trips. Having identified truck IDs, we can now turn to obtain truck trip data between an ID pair. In this paper, we consider trips of truck drivers between a port node and a hinterland node (see Figure 3). Four hinterland nodes, which are strategic in freight flows, are considered at shorter and longer distances

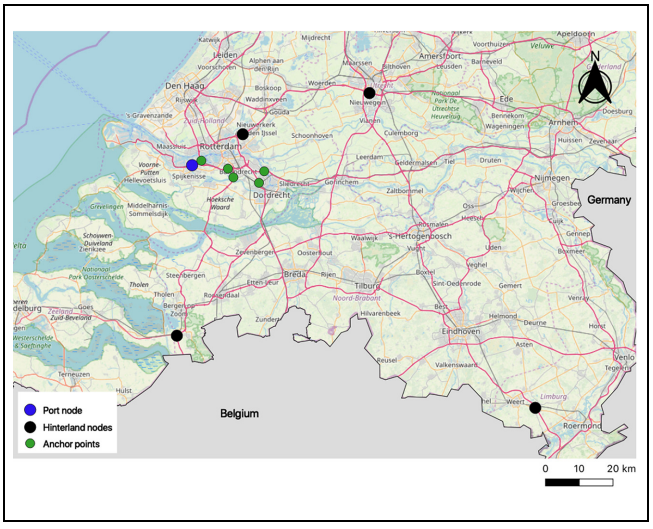

Figure 3. Locations of the port node, hinterlands nodes, and anchor points in the Netherlands.

from the port—a total of eight O-D pairs by considering trips in both directions for a single O-D pair. Since Bluetooth observations lack information about the route chosen by a truck driver between an O-D pair, we use anchor points to alleviate some of the limitations of the Bluetooth data set. An anchor point is defined as a Bluetooth station that lies between an origin and a destination node. Thus, the trips made by truck drivers in our data represent journeys over an origin node, an anchor point, and a destination node.

In addition, we also filter out anomalies (e.g., long breaks) occurring in the trip data using a rule-based approach. Let $TT_{obs,n}$ be the journey time incurred by a truck driver n while making a trip over an origin node, an anchor point (a_n), and a destination node. This travel time is retrieved from the Bluetooth data set. An anchor point allows us to reduce the choice set for a truck driver n, that is, C_n to a viable subset A_n. All those route alternatives for a truck driver n that pass through the anchor point a_n are present in A_n. TT_{in} refers to the expected travel time over a route alternative i for a truck driver n. Then, the journey time ($TT_{obs.n}$) of a truck driver n should lie between the minimum expected travel time and the maximum expected travel time among route alternatives present in the viable choice set A_n. A tolerance of 10% is added to the minimum and maximum expected travel times. We assume that any trip beyond this threshold would have incurred long breaks. Therefore, a continuous trip should satisfy the following Equation 1:

$$0.9 * (\min_{i \in A_n} TT_{in}) \leqslant TT_{obs,n} \leqslant 1.1 * (\max_{i \in A_n} TT_{in}) \quad (1)$$

This step has produced a total of 14,928 trips made by truck drivers during October and November 2017. Next, we present key attributes that characterize a truck trip.

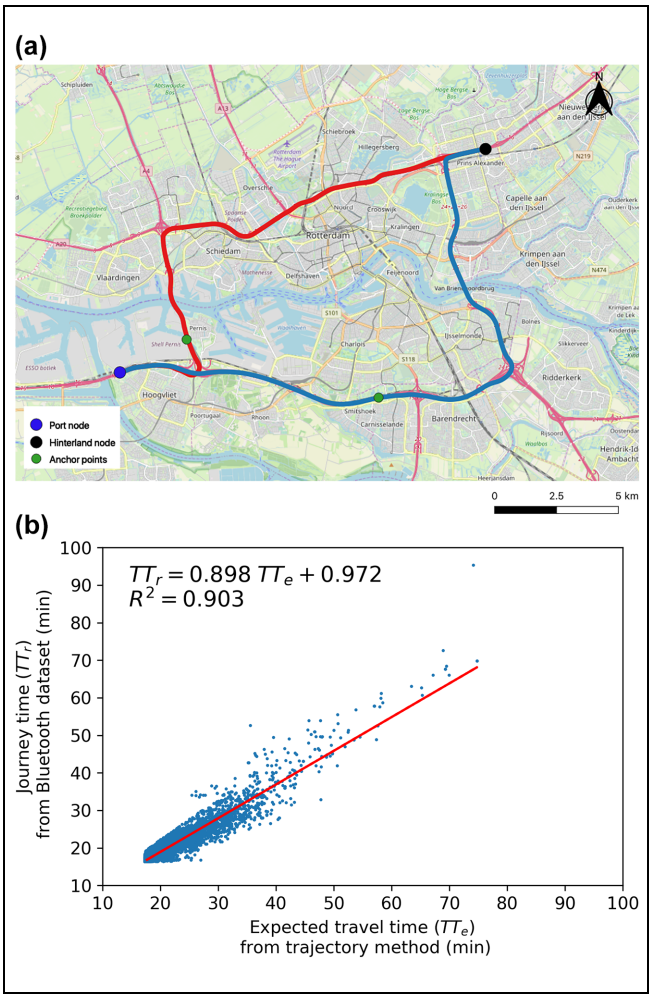

Figure 4. Travel time comparison between the journey time obtained from Bluetooth data and the expected travel time derived from the filtered speed-based (FSB) trajectory method: (*a*) origin–destination pair with two known routes; and (*b*) travel time comparison.

Attributes of Route Alternatives

We consider three attributes: expected travel time, travel distance, and travel time unreliability at the time of departure.

Expected Travel Time. We use expected travel time at the time of departure as one of the attributes of route alternatives. We use loop-detector data (*23*) and apply the filtered speed-based (FSB) trajectory method (*19*) to compute the expected travel time for a truck driver over a path between an O-D pair. In the Netherlands, loop detectors are roughly located at every 500 m and can provide dense coverage of the road network. Between the O-D pair shown in Figure 4, for which route choices are known beforehand, we compared the expected travel time with the journey time obtained from Bluetooth

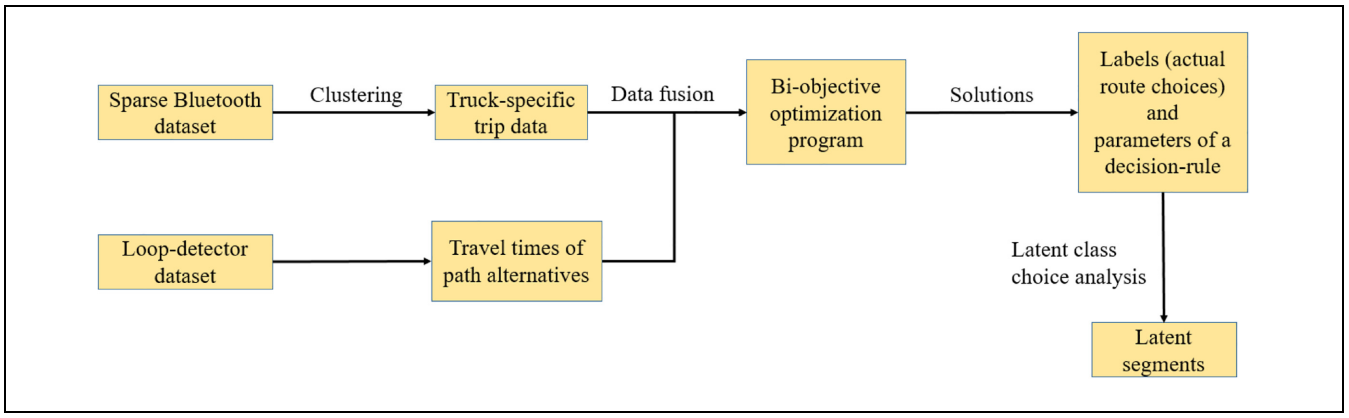

Figure 5. A framework to estimate route choice characteristics of truck drivers from sparse Bluetooth data.

data. The t-test shows that the journey time obtained from the Bluetooth data set (TT_r) and the expected travel time computed from the trajectory method (TT_e) are equal (t-statistic = 11.37, p-value = 8.08e-30). The unit of expected travel time is minutes.

Travel Distance. The travel distance of a route alternative between two Bluetooth stations is measured using Google Maps API. The unit for travel distance is kilometers.

Travel Time Unreliability at the Time of Departure. We use a skewness-based travel time unreliability indicator (*17*). This can be interpreted as the likelihood of incurring a very bad travel time relative to the median travel time, as defined in Equation 2:

$$\lambda_{skew} = \frac{TT_{90} - TT_{50}}{TT_{50} - TT_{10}}, \quad (2)$$

where λ_{skew} is the measure of travel time unreliability and TT_x refers to x percentile of travel time observations. λ_{skew} is computed for the four time periods in a day: morning peak hours (06:30–09:30), day (09:30–16:00), evening peak hours (16:00–19:00), and night (19:00–06:30). Morning and evening peak hours were then combined into peak hours. Day and night constituted off-peak hours. For the computation of travel time unreliability, we select the travel times incurred over a path in the previous 10 days. Having discussed the data set and the attributes, the next section presents our methodology to estimate route choice models using sparse data.

Methods

This paper proposes a new model estimation framework to estimate route choice characteristics from a sparse AVI or Bluetooth data set. Figure 5 presents this framework that accepts truck-specific trip data, obtained through clustering, and the travel times of alternative paths derived from loop detectors as inputs. Subsequently, a bi-objective program is formulated to simultaneously infer actual route choices and the parameters of a route choice model. Finally, a latent class choice analysis is conducted to identify segments with truck drivers' route choice behavior.

The rest of the section is divided into three parts. The first part presents the problem formulation and solution approach. The second part discusses the decision rules that capture the behavior of decision makers. Finally, the third part describes our approach to generating choice sets.

A Bi-Objective Optimization Approach to Simultaneously Infer Actual Route Choices and Estimate the Parameters of a Route Choice Model

Problem Description. This paper proposes a bi-objective model that simultaneously considers the two objectives. On one hand, the proposed model aims to maximize the log-likelihood of an entire data set of choice observations. Here, the likelihood of an entire data set is simply the product of individual choice probabilities. On the other hand, the model aims to minimize the total deviation between the experienced and estimated travel times of a path. The main optimization decisions for the proposed model are as actual route choices (labels) and parameter estimates of a route choice model.

Notations. The mathematical notations used in the paper are listed in Table 1.

Mathematical Model.

$$Max\ F_1 = \sum_{n \in N} \sum_{i \in C_n} y_{in} (\ln P_{in}(\beta)) \quad (3)$$

$$Min\ F_2 = \sum_{n \in N} \sum_{i \in C_n} y_{in} (\ TT_{in} - \ TT_{obs,n})^2 \quad (4)$$

Table 1. Notations

Notation	Description
Indices	
i	Index of a route alternative
n	Index of a truck driver
Sets	
N	Set of truck drivers, $n \in N$ or $N = \{1, ..., n\}$
C_n	Set of route alternatives for a truck driver n
A_n	Set of route alternatives for a truck driver n passing through an anchor point a_n
a_n	Anchor point for a truck driver n
Parameters	
$TT_{obs,\,n}$	Experienced travel time for a truck driver n retrieved from Bluetooth data set
TT_{in}	Estimated travel time for a truck driver n over a route alternative i
β_{min}	The user-specified minimum value for parameters β
β_{max}	The user-specified maximum value for parameters β
Decision variables	
y_{in}	Binary variable, 1 if a truck driver n chooses a route i, 0 otherwise
β	Coefficients of the utility function

Subject to:

$$\sum_{i \in A_n} y_{in} = 1 \ \forall n \in N \tag{5}$$

$$y_{in} \in \{0, 1\} \ \forall n \in N, \ \forall i \in C_n \tag{6}$$

$$\beta_{\min} \leqslant \beta \leqslant \beta_{\max} \tag{7}$$

The objective function (3) maximizes the log-likelihood of the sample. The probability of a truck driver n choosing a route i is expressed by $P_{in}(\beta)$, which depends on the type of decision rule employed. The objective function (4) minimizes the squared deviation between the experienced travel time obtained from the Bluetooth data and the estimated travel time derived from loop-detector data. Constraint (5) ensures that truck drivers can be assigned to, at most, one route that is present in the choice set $A_n \in C_n$. Constraints (6) and (7) state the type of decision variables and their restrictions.

Solution Approach. In bi-objective optimization problems, there is no single optimal solution that can simultaneously optimize all the objective functions. In these cases, decision makers look for the most preferred solution. For these problems, the efficient (or Pareto optimal) solution is the solution that cannot improve one objective function without deteriorating at least one of the rest. A well-known technique for solving bi-objective optimization problems is the ε-constraint method (*24*). This technique optimizes one main objective while other objectives act as constraints. In this paper, our main objective (F_1) is to maximize the log-likelihood of the sample considering F_2 as a constraint (see Equation 8).

$$\begin{aligned} &Max \ F_1 \\ &\text{subject to } F_2 \end{aligned} \tag{8}$$

We consider that F_2 is upper bounded by ε, that is, the total squared deviation between the Bluetooth reported journey time and the expected travel time computed using the trajectory method should not be more than ε. We vary the value of ε from $F_{2,\,\min}$ to $F_{2,\,\max}$ using a payoff table (*24*), which consists of all objective values, when each objective is optimized subject to constraints. The set of all obtained solutions for the entire range of ε are considered the Pareto optimal front of the bi-objective optimization problem. Among the obtained Pareto optimal solutions, the most preferred is selected by the decision maker according to the specific preference of the application. We use an optimization-specific algebraic modeling language AMPL (*25*) to code our optimization formulation and use Bonmin solver (*26*).

Decision Rules

Having formulated the optimization problem formulation, we will now discuss decision rules that describe the process used by the decision maker to choose an alternative. We consider three decision rules: multinomial logit, path size logit, and latent class choice models.

Multinomial Logit Models. Random utility theory assumes that drivers are perfectly rational and that they have perfect discrimination capabilities (*27*). It is assumed that the utility for a driver n associated with route alternative i in the choice set C_n is the sum of a deterministic part (V_{in}) and a random part (ϵ_{in}). We consider a linear utility specification; therefore, we have $V_{in} = \beta X$. Here, β refers to parameters associated with route attributes X. If we assume that the error terms of the utility function are independent and identically Gumbel distributed, the

choice probability of each alternative i can be described in Equation 9 as:

$$P_{in} = \frac{e^{\mu V_{in}}}{\sum_{j \in C_n} e^{\mu V_{jn}}}. \quad (9)$$

Thereby, μ is a positive scale parameter and is related to the Gumble variable.

Path Size Logit Models. Typically, in route choice modeling, the alternatives are often correlated. We, therefore, use a correction factor (*28*). The path size correction factor quantifies the similarity of a route alternative with other route alternatives present in the choice set and its values range from 0 to 1. A distinct route, which is unique and does not overlap with other route alternatives in the choice set, has a path size of 1. Path size correction for a route alternative i corresponding to a truck driver n is defined in Equation 10 as:

$$PS_{in} = \sum_{a \in_i \Gamma_i} \left(\frac{l_a}{L_i}\right) \frac{1}{\sum_{j \in C_n} \delta_{aj}}, \quad (10)$$

where
a is a link in the route alternative i,
Γ_i is the set of links present in the route alternative i,
l_a refers to the length of link a, and
L_i is the length of route alternative i.
$\sum_{j \in C_n} \delta_{aj}$ indicates the total number of route alternatives, present in the choice set of a driver n, sharing link a. By including a path size (PS) correction factor (*28*), we deal with the correlation among route alternatives. Thus, the choice probability of a driver n to choosing a route alternative i is given by Equation 11:

$$P_{in} = \frac{e^{\mu(V_{in} + \ln PS_{in})}}{\sum_{j \in C_n} e^{\mu(V_{jn} + \ln PS_{jn})}}. \quad (11)$$

Latent Class Choice Models. Latent class models are used to capture unobserved heterogeneity in the behavior of truck drivers (*28*). The underlying assumption is that heterogeneity may be produced by taste variations. The latent class model is given by Equation 12:

$$P_{in} = \sum_{s=1}^{S} \pi_{ns} P_{in}(\beta_s), \quad (12)$$

where β_s are class-specific parameter estimates and π_{ns} is the probability that driver n belongs to a segment s and can be given by Equation 13:

$$\pi_{ns} = \frac{\exp(\delta_s \gamma_n)}{\sum_s \exp(\delta_s \gamma_n)}, \quad (13)$$

where δ_s is a class-specific constant to be estimated and γ_n refers to individual-specific socio-economic characteristics.

For model selection, we use the Bayesian information criterion (BIC) (*29*). The BIC value is defined mathematically in Equation 14:

$$BIC = -2 \ln(\mathcal{L}) + K \ln(n), \quad (14)$$

where
$\mathcal{L}$ is the log-likelihood of the model,
K refers to the number of estimable parameters in the model, and
n denotes the number of observations in the data set.
We compute the BIC value for each model under consideration and select the one with the smallest criterion value (*30*).

Choice Set Generation

Analyzing individual decision making requires knowledge not only of what has been chosen, but also of what has not been chosen. Therefore, we require a set of available alternatives (also termed a choice set) that an individual considers during a choice process. We use the Breadth-First Search on Link Elimination (BFS-LE) method to find repeated least-cost paths between an O-D pair (*31*). This algorithm is a link-elimination approach where links of the current least-cost path are removed one at a time to calculate subsequent least-cost paths. We check the commonality of generated least-cost paths and only store unique paths in the route choice set by using travel time as our cost function. A maximum number of 30 unique paths between an O-D pair served as the termination criteria. This value is set as a target choice set size considering the computational tractability of the bi-objective optimization program.

Results

This section first presents the results of an optimization model that is used to simultaneously estimate route choices and parameters of a route choice model. This section then presents segments of truck drivers using the latent class choice model.

Simultaneous Inference of Actual Route Choices and Estimation of Parameters of Route Choice Models

The Pareto curve captures the trade-off between the two conflicting objectives considered in this paper. Figure 6 illustrates the Pareto curves for the multinomial and path size logit models. The solution "a" is obtained when the deviation is minimized, whereas the solution "b" is obtained when the utilities of drivers are maximized.

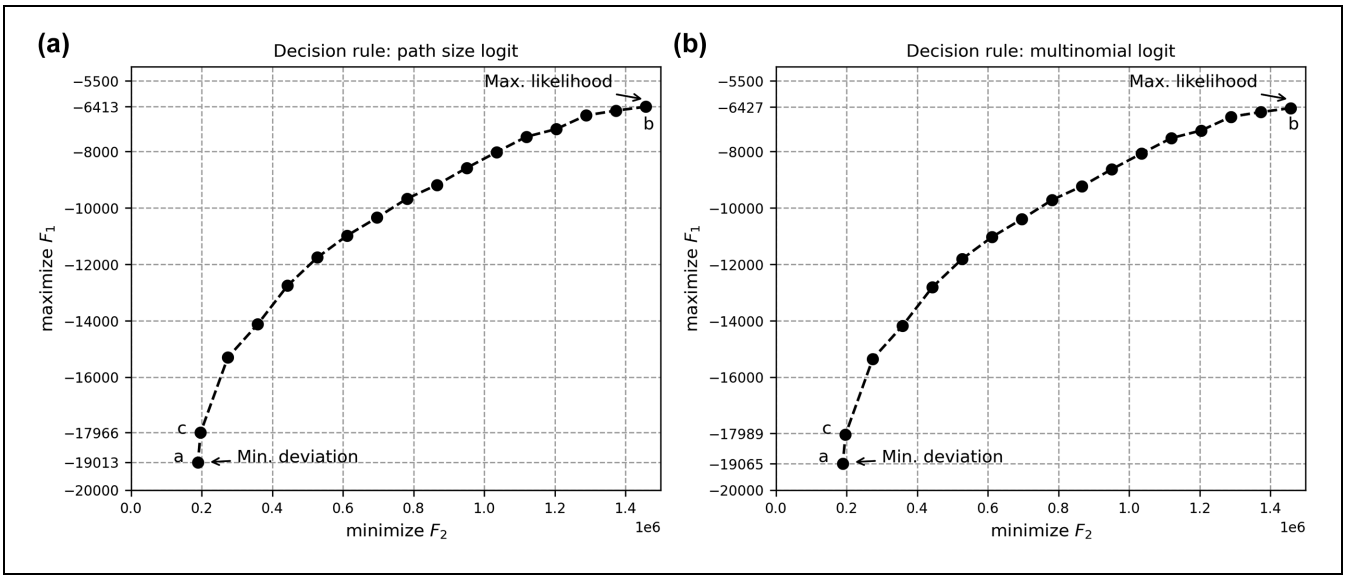

Figure 6. Bi-objective optimal Pareto curve for likelihood and deviation objectives for two different decision rules. The preferred solution is "c" where the deviation is 196,833.64 min^2: (*a*) path size logit model; and (*b*) multinomial logit model.

Among the obtained solutions that lie on the Pareto curve, we select the solution "c" where the value of ε refers to a reasonable estimate of the total squared deviation for sample data. For the O-D pair shown in Figure 4, the mean squared deviation is computed as 5.43 min^2 over an average distance of 31.20 km between the same O-D pair. This average distance is the mathematical average of the length of all route alternatives between an O-D pair. Note that that the mean squared deviation is assumed to increase linearly over longer distances because of errors in loop-detector data or the inability of Bluetooth observations to detect vehicle activity in between. The value of ε for sample data is computed to be 196,833.64 min^2.

Table 2 shows the model fit of the multinomial logit and path size logit models. For both the models, all the parameters are significant and they have expected signs except the travel time unreliability in off-peak hours. The path size logit model outperforms the multinomial logit model based on the likelihood ratio test (p-value < 0.01). The path size logit model not only improves the model fit but can also correct the correlation among route alternatives. The path size logit model shows that truck drivers negatively value travel time, travel distance, and travel time unreliability in peak hours. The path size parameter estimate's positive coefficient is consistent with the findings of Hess et al. (*3*). A positive estimate for path correction term denotes that truck drivers prefer unique routes (i.e., routes with less overlap). We test truck drivers' preferences concerning the travel time unreliability in peak and off-peak hours. During peak hours, truck drivers, in general, stay away from unreliable routes. However, during off-peak hours, they are more likely to make risky route choices. If we now turn to heterogeneity in the route choice behavior of truck drivers, we will present the results of latent class choice models.

Latent Class Choice Models

We use the solutions of the optimization problem generated from the path size logit model as actual route choices of truck drivers to segment the drivers. We estimate a latent class choice model using the PandasBiogeme (*32*). Using the BIC criterion, we find that the model with four segments performs best as it has the least value for the BIC criterion (see Table 3). The proportions of truck drivers belonging to the four-segment model are 15.41%, 37.05%, 39.82%, and 7.72%. The signs of parameter estimates for travel time are negative as expected. Similarly, the signs of parameter estimates for travel distance are negative except for segment 4. The parameters for the route choice model show that truck drivers value the path overlap/correction factor and travel time unreliability differently.

Table 4 shows that segments 2 and 3 constitute three-fourths of the truck drivers. A majority of truck drivers belong to segment 3, and they seem to prefer the shortest distance and shortest time routes. Their preference to choose routes with high overlaps makes them more flexible in unexpected situations such as congestion. However, they show risky behavior during both peak and off-peak hours by having a likelihood for routes with unreliable travel times. Compared with segment 3, truck drivers belonging to segment 2 form a second

Table 2. Route Choice Models for Truck Drivers

Parameters	Path size logit model			Multinomial logit model		
	Value	Rob. SE	Rob. t-test	Value	Rob. SE	Rob. t-test
Natural log of path size based on travel distance	0.492	0.067	7.31	na	na	na
Travel distance (km)	−0.097	0.002	−47.70	−0.090	0.002	−44.00
Expected travel time (min)	−0.023	0.002	−13.30	−0.025	0.001	−13.90
Travel time unreliability if departing in off-peak hours	0.072	0.003	18.40	0.063	0.003	18.00
Travel time unreliability if departing in peak hours	−0.046	0.004	−11.30	−0.035	0.003	−9.72
Number of observations	na	na	14,928	na	na	14,928
Initial log-likelihood	na	na	−25,655.720	na	na	−25,655.272
Final log-likelihood	na	na	−17,966.610	na	na	−17,989.800
Adjusted Rho-square	na	na	0.299	na	na	0.298
Likelihood ratio (LR) test with respect to the multinomial logit model	na	na	46.38	na	na	na
p-value of LR test	na	na	<0.01	na	na	na

Note: SE = standard error; Rob. = Robust; na = not applicable.

Table 3. Comparison of Latent Class Models for Different Number of Segments

Parameters	Latent segments				
	1	2	3	4	5
Final log-likelihood	−17,966.61	−17,054.19	−17,029.96	−16,756.74	−16,860.90
Adjusted Rho-squared	0.299	0.334	0.335	>0.346	0.341
Segment proportions (%)	100.00	37.61	33.91	15.41	1.72
	na	62.39	3.61	37.05	24.55
	na	na	62.48	39.82	10.39
	na	na	na	7.72	43.84
	na	na	na	na	19.50
Estimated parameters	5	11	17	23	29
AIC	35,943.22	34,130.37	34,093.92	33,559.48	33,779.80
BIC	35,981.27	34,214.09	34,223.30	33,734.53	34,000.52

Note: AIC = Akaike information criteria; BIC = Bayesian information criterion; na = not applicable.

Table 4. Segmented Route Choice Model for Truck Drivers

Parameters	Segment 1		Segment 2		Segment 3		Segment 4	
	Value	Rob. SE (t-test statistic)	Value	Rob. SE (t-test statistic)	Value	Rob. SE (t-test statistic)	Value	Rob. SE (t-test statistic)
Class proportion (%)	15.41	na	37.05	na	39.82	na	7.72	na
Natural log of path size based on travel distance	−11.40	3.23 (−3.52)	4.55	0.17 (26.20)	−7.81	0.75 (−10.30)	−40.80	5.18 (−7.88)
Travel distance (km)	−1.31	0.30 (−4.35)	−0.10	0.00 (−30.30)	−0.36	0.02 (−16.50)	0.35	0.05 (6.47)
Expected travel time (min)	−1.90	0.40 (−4.70)	−0.01	0.00 (−5.99)	−0.05	0.01 (−7.04)	−0.05	0.01 (−3.48)
Travel time unreliability if departing in off-peak hours	−0.49	0.10 (−4.67)	0.06	0.01 (11.60)	0.07	0.01 (4.95)	0.28	0.03 (7.17)
Travel time unreliability if departing in peak hours	0.48	0.11 (4.22)	−0.02	0.01 (−2.50)	−0.06	0.02 (4.30)	0.17	0.21 (0.78)
Number of observations	na	na	na	na	na	na	na	14,928
Initial log-likelihood	na	na	na	na	na	na	na	−25,655.27
Final log-likelihood	na	na	na	na	na	na	na	−16,756.74
Adjusted Rho-square	na	na	na	na	na	na	na	0.346

Note: SE = standard error; Rob. = Robust; na = not applicable.

majority and they are less likely to prefer the shortest distance and shortest time routes. Unlike those in segment 3, truck drivers in segment 2 show a preference for routes with less overlap. Their preference to select a unique route is in contrast to the behavior shown by a majority of truck drivers. However, they are concerned about the reliability of travel times during peak hours which prompts them to make informed routing decisions and decreases the possibility of incurring longer travel times. Around a quarter of truck drivers belong to segments 1 and 4. Truck drivers belonging to segment 1 behave more like those in segment 3 except for their sensitivities to the unreliability of travel times. Their sensitivities to travel time and distance are similar to other segments, that is, they negatively value longer time or distance routes. However, they are prone to choose unreliable routes during peak hours. Truck drivers in segment 4 account for only 7.72%. They value shortest time routes but they have an unexpected affinity for longer distance routes. This can be explained by their preference to choose a route with high overlaps. In doing so, they travel longer distances between an O-D pair. Also, they are more likely to choose an unreliable route since they are not significantly affected by the unreliability of travel times during peak hours.

Discussion

This section begins by discussing the plausibility of the route choice characteristics of truck drivers estimated from a sparse Bluetooth data set. We then elaborate on the advantages and limitations of our estimation approach from an application perspective. Finally, we provide the implications of our findings for the design of policies.

This paper found that truck drivers can be segmented into four groups based on their preferences with regard to travel distance, expected travel time to destination, and the unreliability of travel time on a route at the time of departure. The number of segments is greater than used in previous research (*5*, *7*). A possible explanation for this might be that previous research (*5*, *7*) used data from SP surveys, while this paper has used empirical data that include choices made in real-world situations. Another possible explanation is that previous research focused on different business and demographic needs (urban logistics in the Netherlands (*5*) and regional logistics in Washington state, U.S.A. [*7*]) compared with the port logistics considered in our paper. Another significant aspect of our findings is that a majority of truck drivers prefer paths with a high degree of overlap, which indicates that they value the availability of many alternatives to minimize the risks during their trips. These results corroborate the findings of Anderson et al. (*33*). In addition, truck drivers seem to prefer paths with unreliable travel times during peak hours. This outcome is contrary to a previous study that suggested that truck drivers value reliability (*34*). Our results seem to be consistent with Luong et al. (*13*). These results are likely to be related to the behavior of short-haul truck drivers or those departing in the off-peak hours who may take the chance of reaching their destination by choosing an unreliable path.

Let us now turn to our estimation approach, which is formulated as a bi-objective optimization program. Our study raises the possibility that passive data sources (Bluetooth data and loop-detector data) can be used to estimate route choice models. This approach might alleviate the need to perform expansive data collection from SP or RP surveys to understand driving behavior. Different types of fixed-location sensors other than Bluetooth, such as cameras, Wi-Fi sensors, and mobile phone towers can be used as inputs. This formulation can also be applied to freight-specific sparse data sets such as freight trip diaries, which also lack actual route choice observations; with these we could develop advanced commodity-specific route choice models. In this way, the estimation approach opens new possibilities to use sparse data sets in generating insights about the route choice behavior of drivers. The following are the limitations of this approach. First, loop detectors in other regions may not be densely located because of the high costs of installation. We recommend considering the use of other data (e.g., floating car data) in such cases. Second, clustering of vehicles based on speed works well for the Netherlands but may not deliver promising results in countries with low speed limit compliance or different driving policies in place (e.g., a keep-your-lane policy). Here, further research on mode identification from a sparse data set would be recommended.

For the implications for practice, our model indicates that few truck drivers prefer less reliable routes during peak hours. There could be a benefit in including the reliability of travel times in route planning or navigation systems to support companies in making the trade-off between travel time, costs, and reliability. Further research should be undertaken to investigate the objective of truck drivers behind choosing unreliable routes. Also the model could inform the design of interventions by traffic management agencies, such as peak-hour congestion charging or segment-specific route guidance.

Conclusions

This paper estimates the route choice characteristics of truck drivers using sparse automatic vehicle identification (AVI) data. A novel method that uses data fusion and a bi-objective optimization program is proposed to

deal with the sparsity of the AVI data set, which lacks actual route choices (labels). This method can simultaneously estimate the actual route choices and the parameters of a route choice model. This method is successfully applied on a sparse Bluetooth data set of truck drivers making trips to and from the port of Rotterdam in the Netherlands. The resulting models can identify four latent segments within the route choice behavior of truck drivers and capture the effects of time of day (peak and off-peak hours).

In future investigations, it might be possible to incorporate panel effects (or repeated choices made by a driver) within the current framework. Despite the usefulness of our estimation method in delivering behaviorally consistent findings, future work is required to establish the validity of this method. A possible approach would be to conduct a driver survey that can supply the ground truth in addition to a sparse data set. A further study on investigating the route choice behavior of car drivers and its comparison with this study could provide useful insights for the management of significant freight corridors.

Author Contributions

The authors confirm contribution to the paper as follows: study conception and design: Salil Sharma, Hans van Lint, Lóránt Tavasszy, Maaike Snelder; data collection: Salil Sharma; analysis and interpretation of results: Salil Sharma, Hans van Lint, Lóránt Tavasszy, Maaike Snelder; draft manuscript preparation: Salil Sharma, Hans van Lint, Lóránt Tavasszy, Maaike Snelder. All authors reviewed the results and approved the final version of the manuscript.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Netherlands Organization for Scientific Research (NWO), TKI Dinalog, Commit2data, Port of Rotterdam, SmartPort, Portbase, TLN, Deltalinqs, Rijkswaterstaat, and TNO under the project "ToGRIP-Grip on Freight Trips."

ORCID iDs

Salil Sharma https://orcid.org/0000-0002-3646-6940
Lóránt Tavasszy https://orcid.org/0000-0002-5164-2164

References

1. Eurostat. *Freight Transport Statistics - Modal Split*. Statistics-Explained, 2021. https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Freight_transport_statistics_-_modal_split. Accessed June 6, 2021.
2. TLN. Economische Wegwijzer 2020. 2021. https://www.tln.nl/app/uploads/2020/11/TLN_EcoWegwijzer_2020_A4_2P_DEF_RGB_HR.pdf. Accessed January 7, 2021.
3. Hess, S., M. Quddus, N. Rieser-Schüssler, and A. Daly. 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. Arentze, T., T. Feng, H. Timmermans, and J. Robroeks. Context-Dependent Influence of Road Attributes and Pricing Policies on Route Choice Behavior of Truck Drivers: Results of a Conjoint Choice Experiment. *Transportation*, Vol. 39, 2012, pp. 1173-1188.
5. Feng, T., T. Arentze, and H. Timmermans. Capturing Preference Heterogeneity of Truck Drivers' Route Choice Behavior With Context Effects Using a Latent Class Model. *European Journal of Transport and Infrastructure Research*, Vol. 13, 2013, pp. 259-273.
6. Peeta, S., J. L. Ramos, and R. Pasupathy. Content of Variable Message Signs and On-Line Driver Behavior. In *Freight Travel Behavior, Route Choice Behavior, and Advanced Traveler Information Systems: Planning and Administration of Work*, Transportation Research Board, National Research Council, Washington, D.C., 2000.
7. Rowell, M., A. Gagliano, and A. Goodchild. Identifying Truck Route Choice Priorities: The Implications for Travel Models. *Transportation Letters*, Vol. 6, 2014, pp. 98-106.
8. Toledo, T., Y. Sun, K. Rosa, M. Ben-Akiva, K. Flanagan, R. Sanchez, and E. Spissu. Decision-Making Process and Factors Affecting Truck Routing. In *Freight Transport Modelling* (M. Ben-Akiva, H. Meersman, and E. Van de Voorde, eds.), Emerald Group Publishing Limited, Bingley, 2013, pp. 233-249.
9. Knorring, J. H., R. He, and A. L. Kornhauser. Analysis of Route Choice Decisions by Long-Haul Truck Drivers. *Transportation Research Record: Journal of the Transportation Research Board*, 2005. 1923: 46-60.
10. Oka, H., Y. Hagino, T. Kenmochi, R. Tani, R. Nishi, K. Endo, and D. Fukuda. Predicting Travel Pattern Changes of Freight Trucks in the Tokyo Metropolitan Area Based on the Latest Large-Scale Urban Freight Survey and Route Choice Modeling. *Transportation Research Part E: Logistics and Transportation Review*, Vol. 129, 2019, pp. 305-324.
11. Ben-Akiva, M. E., T. Toledo, J. Santos, N. Cox, F. Zhao, Y. J. Lee, and V. Marzano. Freight Data Collection Using GPS and Web-Based Surveys: Insights From US Truck Drivers' Survey and Perspectives for Urban Freight. *Case Studies on Transport Policy*, Vol. 4, 2016, pp. 38-44.
12. Sharma, S., M. Snelder, and H. van Lint. Deriving On-Trip Route Choices of Truck Drivers by Utilizing Bluetooth Data, Loop Detector Data and Variable Message Sign Data. *Proc., 6th International Conference on Models and Technologies for Intelligent Transportation Systems (MT-ITS)*, Cracow, Poland, 2019, pp. 1-8.
13. Luong, T. D., D. Tahlyan, and A. R. Pinjari. Comprehensive Exploratory Analysis of Truck Route Choice Diversity

in Florida. *Transportation Research Record: Journal of the Transportation Research Board*, 2018. 2672: 152-163.
14. Toledo, T., B. Atasoy, P. Jing, J. Ding-Mastera, J. O. Santos, and M. Ben-Akiva. Intercity Truck Route Choices Incorporating Toll Road Alternatives Using Enhanced GPS Data. *Transportmetrica A: Transport Science*, Vol. 16, No. 3, 2020, pp. 654-675.
15. Cao, Q., G. Ren, D. Li, J. Ma, and H. Li. Semi-Supervised Route Choice Modeling With Sparse Automatic Vehicle Identification Data. *Transportation Research Part C: Emerging Technologies*, Vol. 121, 2020, p. 102857.
16. Viti, F., M. Rinaldi, F. Corman, and C. M. J. Tampère. Assessing Partial Observability in Network Sensor Location Problems. *Transportation Research Part B: Methodological*, Vol. 70, 2014, pp. 65-89.
17. van Lint, J. W. C., H. J. van Zuylen, and H. Tu. Travel Time Unreliability on Freeways: Why Measures Based on Variance Tell Only Half the Story. *Transportation Research Part A: Policy and Practice*, Vol. 42, 2008, pp. 258-277.
18. Bierlaire, M., and E. Frejinger. Route Choice Modeling With Network-Free Data. *Transportation Research Part C: Emerging Technologies*, Vol. 16, No. 2, 2008, pp. 187-198.
19. Van Lint, J. W. C. Empirical Evaluation of New Robust Travel Time Estimation Algorithms. *Transportation Research Record: Journal of the Transportation Research Board*, 2010. 2160: 50-59.
20. Namaki Araghi, B., R. Krishnan, and H. Lahrmann. Mode-Specific Travel Time Estimation Using Bluetooth Technology. *Journal of Intelligent Transportation Systems*, Vol. 20, No. 3, 2016, pp. 219-228.
21. Tukey, J. W. Exploratory Data Analysis. In *Addison-Wesley Series in Behavioral Science: Quantitative Methods*, Pearson, London, UK, 1993.
22. Reynolds, D. *Gaussian Mixture Models*. Springer, Boston, MA, 2009.
23. Regiolab-Delft. Regiolab-Delft: A DiTTLab Project. http://www.regiolab-delft.nl/.
24. Aghaei, J., N. Amjady, and H. A. Shayanfar. Multi-Objective Electricity Market Clearing Considering Dynamic Security by Lexicographic Optimization and Augmented Epsilon Constraint Method. *Applied Soft Computing*, Vol. 11, No. 4, 2011, pp. 3846-3858.
25. AMPL. Streamlined Modeling for Real Optimization. https://ampl.com/.
26. Bonmin. *Bonmin (Basic Open-Source Nonlinear Mixed INteger Programming)*. COIN-OR Project. https://projects.coin-or.org/Bonmin.
27. Ben-Akiva, M., and M. Bierlaire. Discrete Choice Models With Applications to Departure Time and Route Choice. In *Handbook of Transportation Science* (R. W. Hall, ed.), Springer, Boston, MA, 2003, pp. 7-37.
28. Ben-Akiva, M., and M. Bierlaire. Discrete Choice Methods and Their Applications to Short Term Travel Decisions. In *Handbook of Transportation Science* (R. W. Hall, ed.), Springer, Boston, MA, 1999, pp. 5-33.
29. Schwarz, G. Estimating the Dimension of a Model. *The Annals of Statistics*, Vol. 6, No. 2, 1978, pp. 461-464.
30. Burnham, K. P., and D. R. Anderson. Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, Vol. 33, No. 2, 2004, pp. 261-304.
31. Rieser-Schüssler, N., M. Balmer, and K. W. Axhausen. Route Choice Sets for Very High-Resolution Data. *Transportmetrica A: Transport Science*, Vol. 9, No. 9, 2013, pp. 825-845.
32. Bierlaire, M. *A Short Introduction to PandasBiogeme*. Technical Report TRANSP-OR 200605. Ecole Polytechnique Fédérale de Lausanne, Switzerland, 2020.
33. Anderson, M. K., O. A. Nielsen, and C. G. Prato. Multimodal Route Choice Models of Public Transport Passengers in the Greater Copenhagen Area. *EURO Journal on Transportation and Logistics*, Vol. 6, No. 3, 2017, pp. 221-245.
34. Bogers, E. A. I., and H. J. Van Zuylen. The Importance of Reliability in Route Choices in Freight Transport for Various Actors on Various Levels. *Proc., European Transport Conference*, Strasbourg, France, 2004, pp. 149-161.