IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 December 2023, accepted 14 December 2023, date of publication 18 December 2023, date of current version 22 December 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3344156

RESEARCH ARTICLE

Truck Rest Stop Imputation From GPS Data: An Interpretable Activity-Based Continuous Hidden Markov Model

MEHDI TAGHAVI[1], ELNAZ IRANNEZHAD[2], AND CARLO G. PRATO[3]

[1]School of Civil Engineering, The University of Queensland, Brisbane, QLD 4072, Australia
[2]Research Centre for Integrated Transport Innovations (rCITI), School of Civil and Environmental Engineering, The University of New South Wales, Sydney, NSW 2052, Australia
[3]School of Civil Engineering, University of Leeds, LS2 9JT Leeds, U.K.

Corresponding author: Elnaz Irannezhad (e.irannezhad@unsw.edu.au)

ABSTRACT The increasing wealth of truck global positioning system (GPS) data has broadened the opportunities for understanding freight logistics activities and enhancing research capabilities to real-world case studies. A very important piece of information from the planning and regulatory perspective concerns the occurrence and location of rest stops. In this study, we propose a data-driven unsupervised machine learning method to impute truck stop events by using a Continuous Hidden Markov Model (CHMM). Specifically, we estimate the joint probability distribution of a mixture of multivariate Gaussian densities, whose parameters depend on the latent states of a Markov chain. Each density represents a cluster of stops that are identified not only from their spatial proximity but also from their temporal proximity as the clustering of the rest stops depends on latent states that are conditional on expected times retrieved from the observed data. In this study, we applied the proposed method to a database containing more than 71 million GPS records of Australian trucks, and we particularly aimed to identify rest stops based on a list of features related to the locations and the load of the trucks. The results showed that the CHMM could account for the location proximity for different activities of truck drivers, and they were validated against complementary data on truck loads and land use by using a stratified sampling technique. Validation results indicated that 94.1% of the rest stops were correctly identified, and highlighted the advantage of the proposed approach without any requirement of labelled data, driver logbook or complimentary survey.

INDEX TERMS GPS truck data mining, trip segmentation, stop identification, Continuous Hidden Markov Model, freight transportation.

I. INTRODUCTION

Agrowing interest has emerged in the ability of disaggregated behavioural models to represent activity patterns and decisions by logistics actors when strategic improvements are made to data collection methods [1], [2], [3]. Over the last two decades, the increasing prevalence of global positioning systems (GPS) devices on trucks has generated a great wealth of mobility-related freight transportation data. These data have broadened the opportunities for understanding freight logistics and enhancing research capabilities to real-world case studies while going well beyond the capabilities of aggregated models.

In terms of freight transport, GPS truck data sources can be categorized into three groups in the literature. The first group comprises data retrieved from GPS units installed on probe trucks participating in pilot projects that typically include a limited number of vehicles and generally are non-random samples. Accordingly, these data do not necessarily provide an unbiased representation of truck travel and trip characteristics (e.g., [4], [5], [6], [7], [8], [9]). The second group includes data obtained from continuous long-term GPS-tracking services that are used for fleet management purposes [10]. These services are growing in the number of subscriptions and this

The associate editor coordinating the review of this manuscript and approving it for publication was Claudio Loconsole.

data collection method is commonly used in urban logistics planning by private transport carriers.

However, these data do not necessarily provide a big-picture perspective as the number of heavy vehicles in urban areas is generally limited, and the limited geographic scope translates into a lack of important information about long-haul freight activities (e.g., trip origin-destination, trip chains). The third group contains data from GPS sources within regulatory and compliance monitoring schemes at the state and national levels (e.g., [11], [12], [13], [14], [15], [16], [17]). These schemes are conducive to the collection of a large amount of data and the implementation of analytical approaches to reveal underlying patterns in freight logistics processes, in addition to seasonal and time-based variations. This study belongs to the third group as we analysed four years of raw GPS data collected from high-performance Australian trucks, totalling more than 71 million records produced as part of the national ''Intelligent Access Program (IAP)'' scheme.

In the realm of extracting insights from passenger GPS data, the existing literature predominantly revolves around the inference of transport modes and activities. Various machine-learning methodologies have been explored for this purpose, including but not limited to random forests [18], support vector machines [19], decision trees [20], and Bayesian networks [21]. Moreover, more sophisticated approaches employing deep learning (DL) techniques have gained prominence, with studies incorporating recurrent and convolutional neural networks (CNN) [22], [23], [24]. When employing DL models with GPS data, a critical preprocessing step involves converting the raw data into trip segments. These segments are then concatenated, each akin to a channel in an RGB image. Subsequently, feature extraction from each segment, including information such as transport mode or activity, is accomplished using a DP framework based on a CNN [25], or a hybrid approach involving clustering methods and CNN [26]. In the latter method, a clustering layer is appended to the embedding layer of the convolutional autoencoder, aiming to strike a balance between the model's reconstruction and clustering losses [26]. However, the task of trip segment identification is inherently challenging due to factors such as GPS measurement errors, diverse traffic conditions, topological variations, and environmental factors [25], [26]. Researchers have addressed these challenges through strategies such as clustering trip segments based on relative distance and speed, treating them as sequences of numerical features [25] or leveraging labels provided by GPS users [11]. The identification of transport modes often relies on point-level motion features like velocity, acceleration, and jerk, followed by the application of unsupervised DL methods to infer transport modes.

In terms of truck GPS data, the literature in freight transportation recognises that a wide variety of types exists in order to capture and understand several aspects of truck movement. Examples include stop and trip-leg identification [4], [27], [28], tour characteristics [7], [17], travel time reliability and terminal performance [12], [29], [30], [31], identification of key supply chain players [32], toll impact on truck-speed profiles [33], and route choice behaviour [15], [34], [35], [36], [37], [38]. Notably, stop identification constitutes a crucial piece of information that is common to several of the aforementioned studies and is positioned prior to identifying trips, configuring routes, and performing other ancillary analyses [39]. The identification of stop locations helps logistics firms, analysts, infrastructure managers and regulators to collect and understand critical information about where and when truck drivers stop to fulfil their legal rest requirements, load or unload their cargo, refuel their vehicle and/or conclude their trip.

Previous studies primarily used a two-step approach in stop identification. The first step usually involved the definition of thresholds for maximum speed and minimum dwell time for stops to be identified within acceptable limits. Moreover, if the duration of the stop event (i.e., the time between the first and last observed points) exceeded a predefined threshold, the sequential GPS points within that duration were considered as a stop or otherwise recognised as ''noise''. For example, these GPS points could reflect stopping at an intersection, dropping suddenly speed because of traffic flow interruptions, or having an error in the GPS readings. Notably, the definition of the dwell time threshold is crucial for stop identification because values that are too long or too short could miss stops and/or incorrectly segment trips. Setting the appropriate dwell time threshold depends on the traffic conditions in a given area [40], and this explains substantial differences observed in the literature, with thresholds equal to 2 [36], 3 [41], 5 [15] and 15 minutes [42].

The second step typically applied a broad range of methods to differentiate trip-end stops from stops caused by congestion or other factors that are less important from the perspective of identifying logistics-related information. The methods include hierarchical clustering [27], Support Vector Machine [28] and the Entropy-based method [42], [43], and all aim to classify stops into (i) primary stops where cargo was loaded or unloaded, and (ii) secondary stops where a truck was stationary for purposes such as breaks, rest, and/or refuelling. Given the legal requirements for truck drivers to take rest, the identification of rest stops is essential for regulatory purposes but is helpful also for planning purposes when modelling route choices [44]. Notably, the significance of rest stops in truck trip chains has been somewhat overlooked [45].

Most analytical approaches to truck stop identification and stop type inference are based on the assumption that a direct relationship exists between a facility location and the type of activity taking place at that location. While associating a location with an activity may be correct for major terminals or large facilities with designated service areas, inaccuracies may arise when analysing a network at a national level. Also, more inaccuracies may emerge when considering that one location might be used for different activities by different truck drivers: for example, gas stations might be used for refuelling by one driver, parking by a second one, eating by a third one, and resting by a fourth one.

To overcome the challenges of the existing truck GPS data mining methods.Taghavi et al. [46] proposed a semi-supervised Hidden Markov Model (HMM) to segment the truck trajectories and to infer three types of stops, namely activity stops, non-activity stops, and stops due to traffic congestion. In that study, 50 long-haul trajectories were chosen as a training set to identify and label trajectory subsections by overlaying the data points with land use data and manual inspections. This manual training was, however, time-consuming and hindered the transferability of the method. Another significant limitation of the previous study consisted in the fact that the proposed semi-supervised method could infer three types of stops, namely (i) activity stops, (ii) non-activity stops, and (iii) stops due to congestion, but was unable to impute rest stops and hence hindered the ability to capture one of the most important pieces of information to understand rest-taking behaviour of truck drivers. Using the same dataset, this current study extends the previous model to infer the rest stops from other activity stops.

In this study, we propose a Continuous Hidden Markov Model (CHMM) to cluster the stop events by type and impute the rest stops for state-wide long-haul truck trips. The application of CHMM in transport data mining was first introduced by Han and Sohn [47] to impute the sequence of trip activities from smart card data in the Seol metropolitan area. The applicability of CHMM in a big and noisy GPS dataset is yet to be tested, particularly for truck trips given their unique specification.

Our approach can address the significant limitations of other methods (as depicted in Table 1) and differentiates from the existing literature in that it captures truck drivers' behaviour and calculates the probability of the activity type in each stop, given the imputed activity type in the previous stop. Moreover, we leverage the huge volume of the observed GPS truck data recorded in the IAP scheme (over 71 million records) to demonstrate that the unsupervised CHMM contributes to the literature by providing the following:

i. independence from the need for labelled data or complementary datasets such as shipper surveys, carrier surveys, or driver logbooks that are not always available, costly to collect, and not transferable; hence adaptability to other datasets with similar characteristics;
ii. ability to learn the continuous progression of different stop events on a given trajectory, despite their irregular duration and ability to incorporate dynamic priors and ability to deal effectively with noise measurements;
iii. ability to account for different activities in one location by various truckers, and inherent support for the abstraction of an activity type via latent states.

Notably, the rich data source used in this study is not limited to a fleet of a single company or a fleet composition of homogeneous vehicles. As a result, this study provides detailed and representative insights into the dynamics of the movement of heterogeneous freight vehicles on a very large-scale road network. We employed the depmixS4 package [48]

TABLE 1. Limitations and challenges of existing models.

Method	Challenges	Advantages of CHMM
Using a secondary dataset to determine the stops (e.g. driver logbook or land use data)	Given the sheer size of the data, manual checks that are widely employed to refine the results of the conventional clustering methods would not be feasible.	CHMM does not require a secondary dataset and is tansferrable to othe cases.
Setting a criterion for the minimum dwell time	The thresholds should vary across urban and non-urban areas. A wide range of threshold values have been considered in the literature (2-15 minutes) based on the traffic condition. The methods is transferrable as do not need adjustments.	CHMM can predict the observation sequences by considering the entire trajectory of a long and diverse truck trip spanning urban and non-urban areas, regardless of the traffic condition.
Density-based clustering methods (e.g. K-mean or DBSCAN)	Grouping data points purely based on the vicinity criterion regardless of the time sequences of traces does not yield accurate results due to GPS signal loss or truck idling. The most common clustering model such as the K-Means method requires pre-defining the number of partitions, which introduces non-transferability issues.	CHMM considers an entire trajectory and the time sequence rather than only the density of data points.
Two-step methods by adjusting the clustering results with land use data	Assuming a direct relationship between facility locations and the types of activity taking place at that location may render inaccuracy in some cases.	CHMM does not rely on land use data for inferring the type of stop, hence relaxing the inaccurate assumption that the type of activity is necessarily tied to the facility type.
DL methods	While DL models demonstrate high efficacy in recognizing complex patterns and dependencies in multimodal data such as images or acoustic signals, they are computationally resource-intensive and often perceived as 'black boxes' with less interpretability compared to classical statistical methods.	CHMM proves more efficient for modelling sequential unimodal numerical data, specifically in capturing temporal dependencies within GPS trajectory data. CHMMs offer transparency and interpretability through the definition of states and transitions, and they can perform well even with limited data. For tasks involving unlabelled trip segment identification, a CHMM may suffice without the need for intricate CNNs or neural networks.

in R for both the estimation and validation of the CHMM in our analysis.

The remainder of the paper is organised as follows. We introduce the formulation of the CHMM in the next section, followed by the description of the data and the implementation details. Then, we present the results and the validation of the model. Lastly, we summarise the most relevant conclusions and reflect on potential applications and extensions.

II. METHODS AND DATA

A. MODELS

Hidden Markov Models (HMMs) are statistical models of systems that are assumed to be Markov processes, where the states are not directly visible, but the outputs (dependent on the states) are visible [49]. In HMMs, the observations in the model are probabilistic functions of the states and the foundation of the model is a doubly embedded stochastic process where the underlying stochastic process is not observable (namely, hidden), but it can be tracked via the other stochastic process that produces the sequence of observations. While HMMs can be used as predictors of observation sequences, in this study an HMM is utilised as an unsupervised learning model to estimate the model parameters from the sequence of observations.

We consider the unknown stop type as a latent state within an N-state space $\{s_1, s_2, \ldots, s_N\}$. As time progresses, a truck either changes its state or remains in its previous state, as shown in Fig. 1. Equation 1 presents the probability π_1 of state s_1 to be the initial state of the trajectory:

$$\pi_1 = P(X_1 = s_1) \tag{1}$$

where X_1 denotes the initial state variable. In an HMM, the movement on the discrete state space is governed by the transition probabilities. The $N \times N$ state transition probability matrix $\boldsymbol{A}$ contains the state transitions $\mathrm{a_{ij}}$ as defined in Equation 2 and shown in Fig. 1:

$$\begin{aligned}\boldsymbol{A} &= \{a_{ij}\}\\ &= P\left(X_t = s_j \,|X_{t-1} = s_i\right), \quad 1 \le i \le N \textit{ and } 1 \le j \le N\end{aligned} \tag{2}$$

where X_t denotes the t^{th} state variable and a_{ij} represents the transition probability that the t^{th} state variable selects stop type j when the previous $(t-1)^{th}$ state was given as stop type i. As will be elucidated in Section III, our focus is on differentiating rest stops from other types of stops through analysis of each state specifications such as duration, transition from other states as well as spot checks. Initially, the optimal number of latent states and their inherent characteristics are unknown. As shown in Fig. 1, only one of the latent states (e.g. State 4) signifies the rest stops while other latent states (State 1 to 3) infer to non-rest stops, associated with either refuelling, congestion or loading/unloading activity stops.

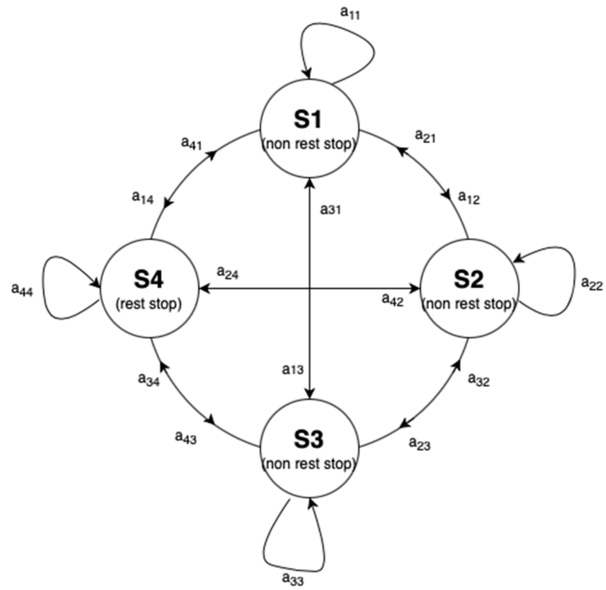

FIGURE 1. Diagram of states and transitional probabilities.

Let $\boldsymbol{O}$ be an observation sequence $\{o_1, o_2, \ldots, o_t, \ldots, o_M\}$, where M is the total number of observations and o_t is an L-dimensional feature vector, namely there are L features that are selected for the preparation of the data, as explained in Section D.

The probability distributions $b_j(o_t)$ of observing o_t at state s_j, known as the "emission probability", composes the $M \times M$ emission probability matrix $\boldsymbol{B}$ in Equation 3:

$$\boldsymbol{B} = \left\{b_j(o_t)\right\} = P\left(o_t \,|X_t = s_j\right), \; 1 \le j \le N \text{ and } 1 \le t \le M \tag{3}$$

In a Continuous HMM (CHMM), both the transitions between hidden states and the arrivals of the observations can occur at arbitrary (continuous) times [50], [51]. Given the problem at hand, where there are irregularly sampled temporal data, it is suitable to consider continuous observations of stop events.

Accordingly, in a CHMM the emission probability $b_j(o_t)$ is presented as a continuous probability density function, generally a mixture of K Gaussian distributions of the observations as presented in Equation 4:

$$\begin{aligned}b_j(o_t) &= P\left(o_t \,|X_t = s_j\right)\\ &= \sum\nolimits_{k=1}^{K} c_{jk} f\left(o_t \,|\, \mu_{jk}, \Sigma_{jk}\right), \; 1 \le j \le N \text{ and } 1 \le t \le M\end{aligned} \tag{4}$$

where c_{jk} is the unknown mixed probability that the observation comes from the k^{th} cluster (i.e., mixture component) when the current state is $\mathrm{s_j}$, and $f\left(o_t \,|\, \mu_{jk}, \Sigma_{jk}\right)$ is the k^{th} mixture component that is used to identify the cluster of o_t with a mean feature vector μ_{jk} and a variance-covariance matrix Σ_{jk}.

The $N \times K$ proportion matrix $\boldsymbol{C}$ for a Gaussian mixture model is defined in Equation 5:

$$\boldsymbol{C} = \{c_{jk}\} = \{P(m_t = k \mid X_t = s_j)\}, \quad j = 1, \ldots, N \qquad (5)$$

where m_t denotes a hidden cluster in the feature space for the t^{th} hidden state of a sequence. We assume that the states share a common set of clusters, and hence $c_k = c_{jk}$, $\mu_k = \mu_{jk}$, and $\Sigma_k = \Sigma_{jk}$.

When considering a sequence of truck rest stops, a state (i.e., a stop type) is determined according to the transition probabilities based on the previous state. Simultaneously, a hidden cluster is assigned to each state according to the determined state and the emission probabilities. Based on the determined cluster, observations are drawn from a Gaussian probability distribution with the mean and variance-covariance of the cluster.

Accordingly, a parameter set $\boldsymbol{\lambda} = \{\pi, \boldsymbol{A}, \boldsymbol{C}, \mu_k, \Sigma_k\}$ is estimated by maximising the probability $P(\boldsymbol{O} \mid \boldsymbol{\lambda})$ of the observation sequence $\boldsymbol{O} = \{o_1, o_2, \ldots, o_t, \ldots, o_M\}$, given the model parameters λ, as written in Equation 6.

$$\lambda^* = \underset{\lambda}{arg\,max}\,\{P(\boldsymbol{O} \mid \boldsymbol{\lambda})\} \qquad (6)$$

The following likelihood function of the observation sequence O of stop types can then be written, given T as the length of the stop sequence for a trip trajectory, as in Equation 7:

$$\begin{aligned} &L(\lambda) \\ &= P(O \mid \lambda) \\ &= \sum\nolimits_{X_{t=1,\ldots,T}} P(o_1, .., o_t, .., o_M | X_1, .., X_T, \lambda) P(X_1, .., X_T | \lambda) \end{aligned} \qquad (7)$$

Given all possible states for X_1, the maximisation of the likelihood function in Equation 7 can be handled only by forward-backward algorithms. The Baum–Welch algorithm [52] is the most common algorithm to estimate HMMs and uses a constrained form of the Expected Maximization (EM) method. The constraints come from the fact that the initial state probabilities and each column of both the transition and emission probability matrix should sum up to 1. In this algorithm, the backward and forward variables $\alpha_t(j)$ and $\beta_t(j)$ are created to compute the probabilities, as shown in Fig. 2.

Algorithm: The Baum-Welch algorithm

Initialize $\lambda_0 \leftarrow (\pi, A, C, \mu, \Sigma)$, for each $o \in O$

For $l = l_0, \ldots, l_{max}$

Do the forward recursion procedure

$\alpha_1(j) = \pi_i b_j(o_1)$

$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(j) a_{ij}] b_j(o_{t+1}) \quad for\ 2 \leq t \leq T$

Do the backward recursion procedure:

$\beta_1(j) = 1$

$\beta_t(j) = \sum_{i=1}^{N} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \quad for\ T-1 \leq t \leq 1$

Do the forward-backward computation:

Calculate the probability $\gamma_t(j,k)$ of o_t being at state s_j at time t with the k^{th} mixture component, where $k \in K$

$\gamma_t(j,k) = \frac{\alpha_t(j)\beta_t(j)}{\sum_{i=1}^{N} \alpha_t(i)\beta_t(i)} \times \frac{c_{jk} C(o_t, \mu_{jk}, \Sigma_{jk})}{\sum_{m=1}^{K}(c_{jm} C(o_t, \mu_{jm}, \Sigma_{jm}))}$

Calculate the probability $\xi_t(i,j,k)$ of being at state s_i and s_j at times t and $t+1$ respectively, with the k^{th} mixture component, where $k \in K$

$\xi_t(i,j,k) = \frac{\alpha_t(j) a_{ij} b_j(O_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(j) a_{ij} b_j(O_{t+1} \beta_t(j))} \times \frac{c_{jk} C(o_t, \mu_{jk}, \Sigma_{jk})}{\sum_{m=1}^{K}(c_{jm} C(o_t, \mu_{jm}, \Sigma_{jm}))}$

Update the parameters of λ :

$\pi_i \leftarrow \gamma_1(j,k) = \frac{\alpha_1(j)\beta_1(j)}{\sum_{i=1}^{N} \alpha_1(i)\beta_1(j)} \times \frac{c_{jk} C(o_t, \mu_{jk}, \Sigma_{jk})}{\sum_{m=1}^{K} c_{jm} C(o_t, \mu_{jm}, \Sigma_{jm})}$

$a_{ij} \leftarrow \frac{\sum_{t=1}^{T-1} \xi_t(i,j,k)}{\sum_{t=1}^{T-1}\sum_{j=1}^{N} \gamma_t(j,k)}$

$c_{jk} \leftarrow \frac{\sum_{t=1}^{T} \gamma_t(j,k)}{\sum_{t=1}^{T}\sum_{k=1}^{K} \gamma_t(j,k)}$

$\mu_{jk} \leftarrow \frac{\sum_{t=1}^{T} \gamma_t(j,k) o_t}{\sum_{t=1}^{T} \gamma_t(j,k)}$

$\Sigma_{jk} \leftarrow \frac{\sum_{t=1}^{T}(\gamma_t(j,k)(o_t - \mu_{jk})(o_t - \mu_{jk})')}{\sum_{t=1}^{T} \gamma_t(j,k)}$

$\bar{\lambda} \leftarrow \underset{\lambda}{arg\,max}\{P[O|\lambda]\}$, such that $P(O|\bar{\lambda}) > P(O|\lambda)$

If ($\frac{P(O|\bar{\lambda}) - P(O|\lambda)}{P(O|\lambda)}$ < threshold) then

Terminate

Else

$\lambda \leftarrow \bar{\lambda}$

End

End

$\lambda^* \leftarrow \{\lambda_l\}_{l=0}^{l_{max}}$

Return λ^*

FIGURE 2. The Baum-Welch algorithm.

B. DATA

The case study focused on data collected as part of the IAP Australian national scheme. The IAP scheme is a voluntary program with the aim of reassuring road authorities that heavy vehicles are operating in compliance with the current legislation. In exchange for compliance, operators receive improved access to particular roads (32). Upon enrolment, operators receive a telematics "black box" to be installed on the heavy vehicles for data collection and monitoring.

The dataset used in this study contains the data collected during the first four years of the scheme (2011–2015), it includes over 71 million GPS records of heavy vehicles, it is stored in a PostgreSQL relational database, and it contains the following variables:

i. vehicle ID, namely an anonymous unique identifier of each heavy vehicle;
ii. longitude and latitude of the vehicle position (recorded every 30 seconds);
iii. timestamp (recorded every 30 seconds);
iv. vehicle load at each timestamp.

C. DATA PREPARATION

We defined a truck trajectory as a finite discrete sequence of snapshots alongside the spatial-temporal position of a unique vehicle. We assumed that the trajectory ends where the vehicle remains in the same position for longer than 8 hours. Using the results of a preliminary study (35), we produced a table of sequential stop-move events with the following attributes:

i. vehicle ID;

ii. trajectory ID: a unique sequential identifier per vehicle and per trajectory;
iii. episode ID: a unique sequential identifier per vehicle, per trajectory, and per episode;
iv. episode type (stop, move, non-activity, and traffic congestion);
v. episode duration, measured in minutes;
vi. incremental distance travelled from the start of the trajectory to the start, measured in kilometres at the start of each episode;
vii. incremental time elapsed since the start of the trajectory, measured in minutes at the start of each episode.

This table is the input to the proposed method that is applied in the current study and, as explained in the following section, allows for the separation between activity stops and rest stops.

D. FEATURE EXTRACTION

We defined a coarse list of measures to explore the possibility of linking the propensity of a driver to stop for rest to features related to the stop locations and travel characteristics of the truck. The list of L features considered in the current study includes:

i. the time spent at each stop location (dwell time);
ii. the dwell time at the previous stop;
iii. the stop/arrival time, measured in minutes from the start of the day;
iv. the distance travelled from the start of the trajectory, measured in kilometres;
v. the distance travelled from the previous stop, measured in kilometres.

E. STRATIFICATION

With the aim of reducing the overall variance in the population, we stratified the trajectory data according to the observation of travel patterns and the diversity of frequently visited locations. The purpose of stratification (or cluster analysis) was to increase the predictability of our algorithm and to distinguish the differences between trajectories with different trip ends and various trip specifications. Accordingly, the CHMM was applied to each stratum in a repeated hold-out procedure. For example, trajectories ending in the Port of Brisbane were stratified into three distinct strata according to their other ends being in (i) the western region, (ii) the northern Brisbane suburbs, and (iii) the northern regions of the state outside the northern Brisbane suburbs. Similarly, trips to or from the ports of Mackay and Gladstone in the north of Queensland were stratified into separate strata according to the location of their other ends. Notably, trajectories ending elsewhere were also grouped in a separate stratum.

III. RESULTS

A. MIXTURE GAUSSIAN MODEL RESULTS

The mixture of Gaussian distributions captured the heterogeneity of different populations, each modelled by one of the homogeneous Gaussian distributions. The number K of Gaussian components of the mixture was required to be determined exogenously while striking a balance between a low number, which would not capture the heterogeneity, and a high number, which could over-fit the data and require needless and endless computation time.

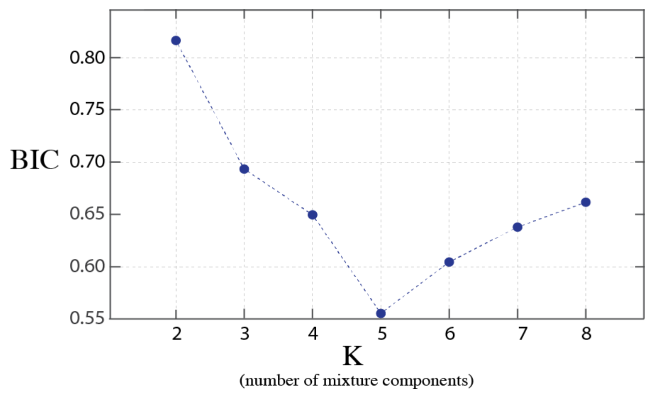

FIGURE 3. BIC values as a function of the number of mixture components.

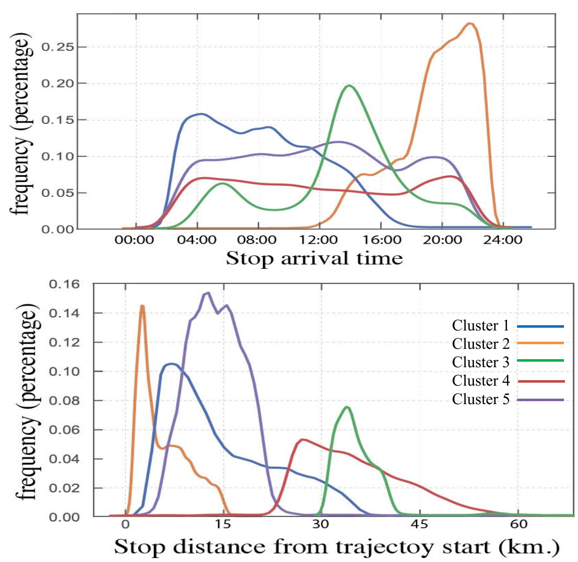

FIGURE 4. Distribution of stop arrival time and stop distance in the 5-cluster solution.

We applied the standard method of estimating mixtures with a variable number of Gaussian components and computing the Bayesian Information Criterion (BIC). Fig. 3 shows the BIC for K varying between 2 and 8, and it is quite clear that the best BIC was obtained for 5 Gaussian components. The interpretation is that there were 5 clusters in the data, and Fig. 4 illustrates the distribution of two selected feature variables across the 5 clusters that were identified by the proposed mixture model.

B. CHMM RESULTS

An iterative approach was used for the algorithm implementation to overcome two possible issues. The first issue concerned the fact that the number of states for the HMM has to be chosen by the analyst, often according to prior knowledge or a specific criterion, since there is no rigorous method to determine the true number of hidden states. The second issue concerned the fact that the EM algorithm utilised for estimating the model parameters might return a local maximum.

Accordingly, for each stratum, we iterated a ten-fold cross-validation process where the number of hidden states increased incrementally from 2 to 10. Within the ten-fold cross-validation process, we employed the depmixS4 package [48] in R to estimate the CHMM parameters with a fixed number of states on a portion of the data, subsequently applying these parameters to the remaining portion for validation. At each run, we obtained reasonable estimates for the five parameter sets: (i) the probabilities of the initial state, (ii) the transition probabilities, (iii) the emission probabilities, (iv) the means of each cluster, and (v) the variance-covariance matrix of each cluster in the feature space.

After the completion of the process, we selected the model with the best goodness of fit over the ten times cross-validation process. It should be noted that it was not possible to compare the likelihood across the nine models, given that the likelihood value was expected to increase with the number of states, and hence we used the BIC values for the selection of the number of clusters. We calculated the normalised BIC value (i.e., the BIC value divided by the total number of predictions) to offer a more direct interpretation when compared to un-normalised likelihood and BIC values.

The interpretation of the latent states requires the assessment of the properties of each state. When observing the frequent transition of state 1 to itself, it is possible to deduce that it does not represent rest stops. Fig. 5(a) and Fig. 5(b) show the dwell time distribution of stops identified as states 2 and 3.

A significant proportion of short dwell time is an indication that these two states might not be rest stops either. Additional spot checks performed on a set of randomly selected stops in these two states found a uniquely high proportion of load transference stops (over 86% of the 150 stops checked manually), which corroborates the fact that they do not represent the rest stops.

State 4, being the remaining state, is most likely to represent the rest stops. Figure 5(c) illustrates the dwell time distribution for stops labelled as state 4. The combination of short and long dwell times in Figure 5(c) strengthens our interpretation of this being the state corresponding to rest stops. Given that the standard working hours of truck drivers are limited by Australian law, we would have expected to see a large proportion of rest stops that take 15 minutes or longer.

Fig. 5(d) shows the frequency of each state along the time elapsed since the start of the trajectory and indicates that the number of stops increases after 5 hours of driving. This is in line with the imposed regulatory scheme requiring drivers to take at least one 15-minute break after five and a half hours of driving. Therefore, we used inductive reasoning based on

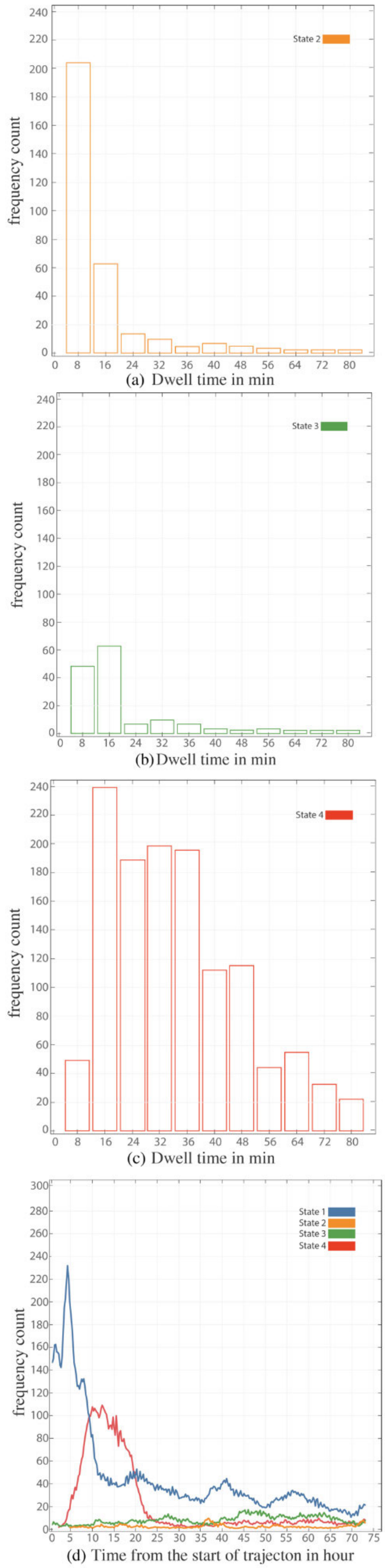

FIGURE 5. Distribution of feature variables for the states.

specifications of diffrernt clusters (dwell time, transitions, spot checks) to make inferences about the nature of different states. This supports the research contribution (ii), identified in Section I.

Similarly, Fig. 5(d) shows that a 15-minute break is the most common duration of rest observed in the data. This observation indicates also that longer rest stops are very likely to be considered by truck drivers who might extend their rest period for a variety of reasons, including avoiding congestion or ensuring that their arrival time at their destination follows a scheduled time.

We used the outcome of the stop type identification model to perform an exemplary secondary analysis of truck movement patterns and long-haul truck network configuration.

Fig. 6(a) provides the spatial distribution of the rest stops (the red dots) with respect to other stop types (the green dots) for a subset of the database. As shown in Fig. 6(b), for example, some stop locations are multipurpose as some truck drivers use these locations as rest stops while others may undertake other types of activities. It should be noted that we do not illustrate incidental stops caused by congestion, signals, or refuelling in this figure.

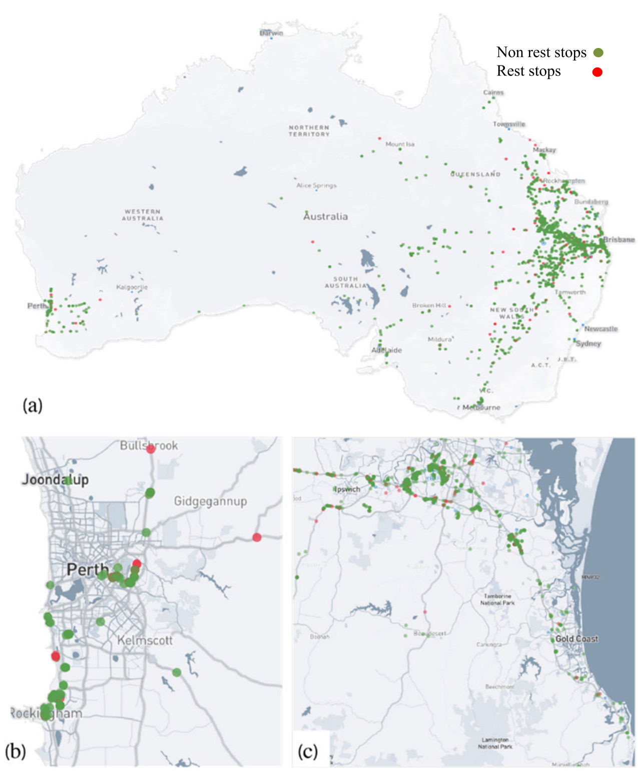

FIGURE 6. Rest stops versus other activity stops.

Fig. 6(b) and Fig. 6(c) present a closer look at two geographically opposite locations, namely southern Queensland on the east coast and Western Australia on the west coast. Rest stops occur close to highways or motorways and are positioned within small developments or outside the boundaries of urban areas. Some facilities seem to attract more truck drivers than other nearby amenities, and many rest stops are co-located with service stations or fast-food plazas. Interestingly, many rest-stop events occur at the same locations where other trucks stop for other activities, which demonstrates a significant feature of our model in that it differentiates multiple activities in the same location. This result supports the research contribution (iii), identified in Section I.

IV. MODEL VALIDATION

Although the proposed methodology adopted an unsupervised machine-learning tool that does not require any calibration process by using labelled data, we performed a validation of our results. In the absence of ground truth, we resorted to using complementary information on truckload and land use to validate empirically the stop types identified by the proposed approach.

We leveraged load data from truck axles, recorded by onboard mass sensors for each GPS data point. Our objective was to discern the purpose of stops, distinguishing between loading/unloading activities or mere rest stops. Additionally, to enhance our comprehension of stops occurring during GPS record interruptions, possibly due to engine shutdown, we cross-referenced information with publicly available sources such as Google Maps. Our analysis involved scrutinizing the surrounding areas of each stop location, identifying indicators of businesses like fuel stations and plazas, as well as assessing spaces conducive to parking, which drivers may use for rest stops. This comprehensive approach facilitated a nuanced understanding of the land use patterns surrounding each stop. Additionally, we consulted the officially designated truck stop areas in Queensland and New South Wales provided by the respective Road authorities to augment our dataset and ensure a comprehensive analysis.

Using the stratified sampling technique, we identified a total of 30 designated truck rest stops in 11 regions. Within each of these sites, we manually examined 3222 trajectories where vehicles had stopped at any of the selected designated rest stops. During this manual examination, we labelled stops that appeared to be rest stops based on their duration stop, travel patterns observed from the start of the trajectory and whether they had rest stops earlier along their trajectory. The labelled outputs from this manual labelling process were then compared against the results generated by our model.

Table 2 provides the validation results. The figures of True Positive (TP) refer to the cases where the model correctly identified a location as a rest stop while False Positive (FP) figures refer to locations that are not designated for a rest stop, however, the model identified them as rest stop.

However, it is uncertain if the model incorrectly identified these locations or if truck drivers took a rest in non-designated areas. True Negative (TN) figures refer to the cases where the model could correctly exclude locations that are not rest stops while False Negative (FN) figures refer to cases where the model fails to identify the designated rest area as a location where drivers of these trajectories have taken rest. It should be again noted that False Negative can be due to undertaking other types of activities such as fuelling instead of resting.

TABLE 2. Validation results.

State	Locality	# of rest sites	TP, FP, FN, TN	Prediction Accuracy (%)	Precision	Recall	F1 Score
Queensland	Emerald	10	159,5,7,17	93.60	0.97	0.95	0.96
	Rockhampton	6	234,8,12,5	92.28	0.96	0.95	0.95
	Marmor	1	9,4,2,3	66.67	0.69	0.81	0.75
	Gatton	1	121,8,8,4	88.65	0.93	0.93	0.93
	Warrill View	1	12,3,1,2	77.78	0.80	0.92	0.85
	Toowoomba	2	130,8,5,12	91.61	0.94	0.96	0.95
New South Wales	Greater argyle	1	642,18,5,18	96.63	0.97	0.99	0.98
	Kempsey	1	302,14,9,6	93.05	0.95	0.97	0.96
	Blue mountains	2	120,9,12,18	86.79	0.93	0.90	0.92
	Wollongong	2	365,18,4,4	94.37	0.95	0.98	0.97
	Wollondilly	3	760,22,8,89	96.59	0.97	0.99	0.98
Total		30	2854,117,73,178	94.10	0.96	0.97	0.96

Hence, these figures can only be truly assessed in the presence of a driver logbook.

As shown in Table 2, the prediction accuracy measure is 94.1% which is the overall correctness of predictions and suggests that 94.1% of the rest stops were correctly matched with the manually constructed stop types, and the majority of the (few) inaccurate matches were short-duration stops. The F1 Score is obtained as 0.968 which identifies the model's ability to find all positive instances of rest stops. The evaluation of the model's performance demonstrates its effectiveness in handling truck GPS noises without relying on labelled data or secondary dataset, hence adaptability to other truck GPS datasets with similar specifications, as delineated in Section I under research contributions (i) and (ii).

V. CONCLUSION

Although stop identification is recognised as an essential element to retrieve a significant piece of information related to freight logistics processes in the existing literature on GPS truck data mining, the identification process is often either overlooked or loaded with significant requirements in terms of labelled data. Given the importance of stop identification in the monitoring efforts for regulatory compliances, the design of policies, and the planning of infrastructure, we proposed a CHMM to identify truck rest stops by exploiting the inherent temporal correlation of the data.

The novelty and key takeaways of this research is as follows. First, the proposed method does not require any other information than the one provided by GPS traces. Second, the proposed method was able to handle over 71 million GPS records with a relatively simple statistical model. When considering the easy implementation of CHMM and the nature of the GPS data, the model appears flexible, and it is likely to be transferable to other datasets. Third, the proposed CHMM provides a powerful tool to extract truck drivers' trip information and the approach proved to be particularly suitable for state-wide long-haul freight trips where the correlation structure of the data can be decomposed according to a finite number of easily interpretable distributions. Fourth, not only does the model extend prior attempts to stop type identification by isolating rest stops from other activity stops, but also it accommodates one location to be used for different activities by truck drivers, a significant departure from more ad hoc methods in the existing literature. Lastly, the most important advantage of the proposed approach is that it is an unsupervised machine-learning method that does not require labelled data, driver logbooks or complementary surveys.

The key pragmatic takeaways can be summarised as follows: (i) a large proportion of rest stops takes 15 minutes or longer, (ii) the number of stops increases after 5 hours of driving, (iii) a 15-minute break is the most common duration of rest observed in the data, (iv) some facilities can attract more truck drivers than other nearby amenities, (v) many rest stops are co-located with service stations or fast-food plazas, which highlights the necessity of relaxing the assumption of the landuse type and type of stop; (vi) many rest stop events occur at the same locations where other trucks stop for other activities, which demonstrates a significant feature of our model in that it differentiates for multiple activities in the same location, and finally (vii) longer rest stops are very likely to be considered by truck drivers who might extend their rest period for a variety of reasons, including avoiding congestion or ensuring that their arrival time at destination follows a scheduled time.

Furthermore, this approach can easily lend itself to more complex models. For example, the model can be adapted to estimate destination choice as well as stop type. The CHMM also allows the inclusion of additional covariates (e.g., spatial relationships) that help predict truck drivers' behaviour while incorporating the temporal correlation. Another reasonably straightforward extension to the model is the inclusion of different information or data to help predict truck drivers' activities (e.g., weather, commodity type) by incorporating the additional information as a covariate in any of the parameters.

Our validation approach rigorously filtered the data to identify designated rest stops by employing various criteria such as duration of stay, travel time since the start of the trajectory, and the presence of other long stops from the start. However, it is important to note that the observed stops used for validation were inferred rather than directly observed, but we ensured that this inference process was conducted independently of the model to maintain objectivity. To further improve the validation process, we recognize the importance of examining the model's performance using a secondary dataset that offers more reliable information on the rest stops. By incorporating such a dataset in future research, the robustness and reliability of validation can be enhanced, thereby establishing a more solid foundation for evaluating the effectiveness of the proposed model.

ACKNOWLEDGMENT
The authors would like to thank the Queensland Department of Transport and Main Roads and the IAP for access to the data used in this study.

REFERENCES
[1] J. Y. J. Chow, C. H. Yang, and A. C. Regan, "State-of–the art of freight forecast modeling: Lessons learned and the road ahead," *Transportation*, vol. 37, no. 6, pp. 1011–1030, Nov. 2010.
[2] D. A. Hensher and S. M. Puckett, "Refocusing the modelling of freight distribution: Development of an economic-based framework to evaluate supply chain behaviour in response to congestion charging," *Transportation*, vol. 32, no. 6, pp. 573–602, Nov. 2005, doi: 10.1007/s11116-004-7615-6.
[3] G. D. Jong, L. Tavasszy, J. Bates, S. E. Grønland, S. Huber, O. Kleven, P. Lange, O. Ottemöller, and N. Schmorak, "The issues in modelling freight transport at the national level," *Case Stud. Transp. Policy*, vol. 4, no. 1, pp. 13–21, Mar. 2016.
[4] S. P. Greaves and M. A. Figliozzi, "Collecting commercial vehicle tour data with passive global positioning system technology: Issues and potential applications," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2049, no. 1, pp. 158–166, Jan. 2008.
[5] H.-C. Chu, "An empirical study to determine freight travel time at a major port," *Transp. Planning Technol.*, vol. 34, no. 3, pp. 277–295, Apr. 2011.
[6] J. Comendador, M. E. López-Lambas, and A. Monzón, "A GPS analysis for urban freight distribution," *Proc. Social Behav. Sci.*, vol. 39, pp. 521–533, Jan. 2012.
[7] K. Sturm, Z. Pourabdollahi, A. K. Mohammadian, and K. Kawamura, "GPS and driver log–based survey of grocery trucks in Chicago, Illinois," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2410, no. 1, pp. 31–38, Jan. 2014.
[8] A. R. Alho, L. You, F. Lu, L. Cheah, F. Zhao, and M. Ben-Akiva, "Next-generation freight vehicle surveys: Supplementing truck GPS tracking with a driver activity survey," presented at the *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 2974–2979.
[9] H. Yang, P. J. Jin, B. Ran, D. Yang, Z. Duan, and L. He, "Freeway traffic state estimation: A Lagrangian-space Kalman filter approach," *J. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 525–540, Nov. 2019.
[10] X. Ma, Y. Wang, E. McCormack, and Y. Wang, "Understanding freight trip-chaining behavior using a spatial data-mining approach with GPS data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2596, no. 1, pp. 44–54, Jan. 2016.
[11] C.-F. Liao, "Generating reliable freight performance measures with truck GPS data: Case study in twin cities metropolitan area, Minnesota," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2410, no. 1, pp. 21–30, Jan. 2014.
[12] M. Flaskou, M. A. Dulebenets, M. M. Golias, S. Mishra, and R. M. Rock, "Analysis of freight corridors using GPS data on trucks," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2478, no. 1, pp. 113–122, Jan. 2015.
[13] A. Thakur, A. R. Pinjari, A. B. Zanjani, J. Short, V. Mysore, and S. F. Tabatabaee, "Development of algorithms to convert large streams of truck GPS data into truck trips," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2529, no. 1, pp. 66–73, Jan. 2015.
[14] A. B. Zanjani, A. R. Pinjari, M. Kamali, A. Thakur, J. Short, V. Mysore, and S. F. Tabatabaee, "Estimation of statewide origin–destination truck flows from large streams of GPS data: Application for Florida statewide model," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2494, no. 1, pp. 87–96, Jan. 2015.
[15] M. Kamali, A. Ermagun, K. Viswanathan, and A. R. Pinjari, "Deriving truck route choice from large GPS data streams," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2563, no. 1, pp. 62–70, Jan. 2016.
[16] N. Chankaew, A. Sumalee, S. Treerapot, T. Threepak, H. W. Ho, and W. H. K. Lam, "Freight traffic analytics from national truck GPS data in Thailand," *Transp. Res. Proc.*, vol. 34, pp. 123–130, Jan. 2018.
[17] S. I. You and S. G. Ritchie, "A GPS data processing framework for analysis of drayage truck tours," *KSCE J. Civil Eng.*, vol. 22, no. 4, pp. 1454–1465, Apr. 2018.
[18] L. Stenneth, O. Wolfson, P. S. Yu, and B. Xu, "Transportation mode detection using mobile phones and GIS information," in *Proc. 19th ACM SIGSPATIAL Int. Conf. Adv. Geographic Inf. Syst.*, Nov. 2011, pp. 54–63.
[19] Z. Sun and X. Ban, "Vehicle classification using GPS data," *Transp. Res. C, Emerg. Technol.*, vol. 37, pp. 102–117, Dec. 2013.
[20] Y. Zheng, L. Liu, L. Wang, and X. Xie, "Learning transportation mode from raw GPS data for geographic applications on the web," in *Proc. 17th Int. Conf. World Wide Web*, Apr. 2008, pp. 247–256.
[21] G. Xiao, Z. Juan, and C. Zhang, "Travel mode detection based on GPS track data and Bayesian networks," *Comput., Environ. Urban Syst.*, vol. 54, pp. 14–22, Nov. 2015.
[22] H. Wang, G. Liu, J. Duan, and L. Zhang, "Detecting transportation modes using deep neural network," *IEICE Trans. Inf. Syst.*, vol. E100.D, no. 5, pp. 1132–1135, 2017.
[23] M. Simoncini, L. Taccari, F. Sambo, L. Bravi, S. Salti, and A. Lori, "Vehicle classification from low-frequency GPS data with recurrent neural networks," *Transp. Res. C, Emerg. Technol.*, vol. 91, pp. 176–191, Jun. 2018.
[24] R. Zhang, P. Xie, C. Wang, G. Liu, and S. Wan, "Classifying transportation mode and speed from trajectory data via deep multi-scale learning," *Comput. Netw.*, vol. 162, Oct. 2019, Art. no. 106861.
[25] S. Dabiri, C.-T. Lu, K. Heaslip, and C. K. Reddy, "Semi-supervised deep learning approach for transportation mode identification using GPS trajectory data," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 5, pp. 1010–1023, May 2020.
[26] C. Markos and J. J. Q. Yu, "Unsupervised deep learning for GPS-based transportation mode identification," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–6.
[27] B. W. Sharman and M. J. Roorda, "Analysis of freight global positioning system data: Clustering approach for identifying trip destinations," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2246, no. 1, pp. 83–91, Jan. 2011.
[28] X. Yang, Z. Sun, X. J. Ban, and J. Holguín-Veras, "Urban freight delivery stop identification with GPS data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2411, no. 1, pp. 55–61, Jan. 2014.
[29] E. McCormack, E. Scharnhorst, W. Zhao, and D. Tabat, *Using GPS Truck Data to Identify and Rank Bottlenecks in Washington State*. Washington, DC, USA: Washington State Transportation Center (TRAC), 2011.
[30] B. W. Sharman and M. J. Roorda, "Multilevel modelling of commercial vehicle inter-arrival duration using GPS data," *Transp. Res. E, Logistics Transp. Rev.*, vol. 56, pp. 94–107, Sep. 2013.
[31] W. Zhao, E. McCormack, D. J. Dailey, and E. Scharnhorst, "Using truck probe GPS data to identify and rank roadway bottlenecks," *J. Transp. Eng.*, vol. 139, no. 1, pp. 1–7, Jan. 2013.
[32] J. W. Joubert and K. W. Axhausen, "A complex network approach to understand commercial vehicle movement," *Transportation*, vol. 40, no. 3, pp. 729–750, May 2013.
[33] Z. Wang and A. V. Goodchild, "GPS data analysis of the impact of tolling on truck speed and routing: Case study in seattle, Washington," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2411, no. 1, pp. 112–119, Jan. 2014.
[34] J. H. Knorring, R. He, and A. L. Kornhauser, "Analysis of route choice decisions by long-haul truck drivers," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1923, no. 1, pp. 46–60, Jan. 2005.
[35] T. Yokota and D. Tamagawa, "Route identification of freight Vehicle's tour using GPS probe data and its application to evaluation of on and off ramp usage of expressways," *Proc. Social Behav. Sci.*, vol. 39, pp. 255–266, Jan. 2012.
[36] S. Hess, M. Quddus, N. Rieser-Schüssler, and A. Daly, "Developing advanced route choice models for heavy goods vehicles using GPS data," *Transp. Res. E, Logistics Transp. Rev.*, vol. 77, pp. 29–44, May 2015.
[37] D. Tahlyan and A. R. Pinjari, "Performance evaluation of choice set generation algorithms for analyzing truck route choice: Insights from spatial aggregation for the breadth first search link elimination (BFS-LE) algorithm," *Transportmetrica A, Transp. Sci.*, vol. 16, no. 3, pp. 1030–1061, Jan. 2020.
[38] T. Toledo, 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, Transp. Sci.*, vol. 16, no. 3, pp. 654–675, Jan. 2020.
[39] L. Gong, T. Morikawa, T. Yamamoto, and H. Sato, "Deriving personal trip data from GPS data: A literature review on the existing methodologies," *Proc. Social Behav. Sci.*, vol. 138, pp. 557–565, Jul. 2014.

[40] J. Du and L. Aultman-Hall, "Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues," *Transp. Res. A, Policy Pract.*, vol. 41, no. 3, pp. 220–232, Mar. 2007.
[41] X. Ma, E. D. McCormack, and Y. Wang, "Processing commercial global positioning system data to develop a web-based truck performance measures program," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2246, no. 1, pp. 92–100, Jan. 2011.
[42] K. Gingerich, H. Maoh, and W. Anderson, "Classifying the purpose of stopped truck events: An application of entropy to GPS data," *Transp. Res. C, Emerg. Technol.*, vol. 64, pp. 17–27, Mar. 2016.
[43] H. Zhao, D. Qian, Y. Lv, B. Zhang, and R. Liang, "Development of a global positioning system data-based trip-purpose inference method for hazardous materials transportation management," *J. Intell. Transp. Syst.*, vol. 24, no. 1, pp. 24–39, Jan. 2020.
[44] M. Rowell, A. Gagliano, and A. Goodchild, "Identifying truck route choice priorities: The implications for travel models," *Transp. Lett.*, vol. 6, no. 2, pp. 98–106, Apr. 2014.
[45] K. Haque, S. Mishra, R. Paleti, M. M. Golias, A. A. Sarker, and K. Pujats, "Truck parking utilization analysis using GPS data," *J. Transp. Eng., A, Syst.*, vol. 143, no. 9, Sep. 2017, Art. no. 04017045.
[46] M. Taghavi, E. Irannezhad, and C. G. Prato, "Identifying truck stops from a large stream of GPS data via a hidden Markov chain model," in *Proc. IEEE Intell. Transp. Syst. Conf. (ITSC)*, Oct. 2019, pp. 2265–2271.
[47] G. Han and K. Sohn, "Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model," *Transp. Res. B, Methodol.*, vol. 83, pp. 121–135, Jan. 2016.
[48] I. Visser and M. Speekenbrink, "DepmixS4: AnRPackage for hidden Markov models," *J. Stat. Softw.*, vol. 36, no. 7, pp. 1–21, 2010.
[49] L. Rabiner and B. Juang, "An introduction to hidden Markov models," *IEEE ASSP Mag.*, vol. 3, no. 1, pp. 4–16, Jan. 1986.
[50] C. H. Jackson, "Multi-state models for panel data: ThemsmPackage for R," *J. Stat. Softw.*, vol. 38, no. 8, pp. 1–28, 2011.
[51] D. R. Cox and H. D. Miller, *The Theory of Stochastic Processes*. Boca Raton, FL, USA: CRC Press, 1977.
[52] L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *Ann. Math. Statist.*, vol. 41, no. 1, pp. 164–171, Feb. 1970.

MEHDI TAGHAVI is currently pursuing the Ph.D. degree with The University of Queensland. He is also a Senior Transport Modeler with The University of Queensland. He has been particularly involved in various strategic transport modeling projects across Queensland.

ELNAZ (ELLI) IRANNEZHAD received the Ph.D. degree in transport engineering from the School of Civil Engineering, The University of Queensland. She is a Senior Lecturer of transportation engineering with the Research Centre for Integrated Transport Innovation (rCITI), School of Civil and Environmental Engineering, The University of New South Wales (UNSW). Prior to joining UNSW, she was employed on a three-year postdoctoral research fellowship with the University of Queensland's School of Economics. She was with the Australian Road Research Board, as a Principal Engineer, for two years. Her research contributes to the advancement of science in cross-disciplinary fields, including logistics, freight transportation, agent-based modeling, automated vehicles, and blockchain technology.

CARLO G. PRATO received the Ph.D. degree from Politecnico di Torino, Italy, where he approached the study of travel behavior by focusing on route choices of car drivers. He is the Head of the School of Civil Engineering and a Professor of transportation engineering with the University of Leeds. Prior to joining the University of Leeds in September 2021, he was with the Technion–Israel Institute of Technology, the Technical University of Denmark, and The University of Queensland.