Journal of Big Data Analytics in Transportation (2019) 1:175–190
https://doi.org/10.1007/s42421-019-00011-x

ORIGINAL PAPER

Exploring Algorithms for Revealing Freight Vehicle Tours, Tour-Types, and Tour-Chain-Types from GPS Vehicle Traces and Stop Activity Data

André Romano Alho[1] · Takanori Sakai[1] · Ming Hong Chua[1] · Kyungsoo Jeong[2] · Peiyu Jing[3] · Moshe Ben-Akiva[3]

Received: 12 March 2019 / Revised: 30 September 2019 / Accepted: 15 October 2019 / Published online: 7 November 2019

Abstract
Freight vehicle tours and tour-chains are essential elements of state-the-art agent-based urban freight simulations as well as key units to analyse freight vehicle demand. GPS traces are typically used to extract vehicle tours and tour-chains and became available in a large scale to, for example, fleet management firms. While methods to process this data with the objective of analysing and modelling tour-based freight vehicle operations have been proposed, they were not fully explored with regard to the implication of underlying assumptions. In this context, we test different algorithms of stop-to-tour assignment, tour-type and tour-chain identification, aiming to expose their implications. Specifically, we compare the traditional stop-to-tour assignment algorithm using the location of a "base" as the start/end point of tours, against other algorithms using stop activities or payload capacity usage. Furthermore, we explore high-resolution tour-type/chain identification algorithms, considering stop types and recurrence of visits. For tour-chain identification, we explore two algorithms: one defines the day-level tour-chain-type based on the predominant tour-type identified for the period of 1 day and another defines the tour-chain-type based on the average number of stops per tour by stop type. For a demonstration purpose, we apply the methods to data from a large-scale GPS-based survey conducted during 2017–2019 in Singapore. We compare the algorithms in an assessment of freight vehicle operations day-to-day pattern homogeneity. Our analysis demonstrates that the predictions of tours, tourtypes, and tour-chain-types are highly dependent on the assumptions used, underlining the importance of carefully selecting and disclosing the methods for data processing. Finally, the exploration of day-to-day pattern homogeneity reveals operational differences across vehicle types and industries.

Keywords Freight flows · Big data analytics · Truck GPS data · Commercial vehicle · Tour analysis

Introduction

GPS data for freight vehicles is increasingly available, due to the deployment of telematics to companies with sizeable fleets and fleet management firms. A widely known example of such data is the American Transportation Research Institute truck GPS dataset (Short 2014). According to some criteria, such as its large volume and by-product nature, GPS traces of freight vehicles can be considered as big data. Dedicated data collection efforts are also becoming more sophisticated, such as the integration of GPS-enabled devices and GPS-enabled digital freight surveys (Alho et al. 2018). As

✉ André Romano Alho
andre.romano@smart.mit.edu

✉ Ming Hong Chua
rayden.chua@smart.mit.edu

Takanori Sakai
takanori@smart.mit.edu

Kyungsoo Jeong
kyungsoo.jeong@nrel.gov

Peiyu Jing
peiyu@mit.edu

Moshe Ben-Akiva
mba@mit.edu

1 Singapore-MIT Alliance for Research and Technology, 1 CREATE Way, #09-01/02 CREATE Tower, Singapore 138602, Singapore

2 National Renewable Energy Laboratory, 15013 Denver West Parkway, Golden, CO 80401, USA

3 Intelligent Transportation Systems Lab, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Room 1-181, Cambridge, MA 02139, USA

a result, more vehicle trajectory and stop-level data is available to further study freight vehicle movements.

However, the methods to process and analyse such data for their use in freight transportation modelling have not been fully explored. Particularly, there is a research gap regarding the conversion of freight vehicle GPS traces into tour-level data. The importance of this process is evident from the fact tours are one of the adopted units for vehicle flow analysis, and furthermore, freight vehicle tours and tour-chains are an integral element of state-the-art agent-based urban freight simulations (Hunt and Stefan 2007; de Bok and Tavasszy 2018). Well-structured and information-rich records of truck tours have the potential to enhance the replication of freight vehicle tour-chains in a simulation environment for policy analysis. The definition of methods for identifying tours directly contributes to tour-chain modelling (Jing et al. 2019), tour-based simulation case studies (Alho et al. 2019, Gopalakrishnan et al. 2019), and the identification of commodity flows and load factors (Alho et al. 2018).

Generally speaking, freight vehicle tours are more challenging to predict than passenger tours. For passengers, *home* and *work* are pivotal points around which tours and sub-tours occur. On the other hand, a single freight vehicle might visit multiple overnight parking locations (Alho et al. 2018), which results in tour-chains having different start/end points at a daily level. Several other challenges are detailed by You et al. (2016), such as the limited data availability and increasing trip chaining behaviour (comparatively to passenger tours). It must be acknowledged that, ideally, data on the "ground truth" regarding stop-to-tour membership, tour-type and tour-chain would be collected. To the best of our knowledge, there is no consensus on the definition of a "freight vehicle tour". In other words, the criteria which define the start and end of a tour are not well-established, which further justifies the research in this paper. This research sets to explore the output differences arising from the various assumptions in the stop-to-tour assignment process as well as in the tour-type and tour-chain identification processes. A descriptive analysis follows, to illustrate such implications, where an application is focused on the prediction of day-to-day pattern homogeneity and the differences across sub-populations. Follow ups to the analysis in this paper include the exploration of concepts such as tour typology by, for example, the characteristics of operator, commodity handled, vehicle type and transportation service, and tour topology (e.g., spatial tour characteristics such as spatial coverage and displacement) by subpopulation.

The rest of this paper is organised as follows. The second section provides a literature review, covering the definitions of a set of terms that are still not standardised in this knowledge domain; the third section describes a description of the data requirements for the analyses performed in this paper as well as the selected sample; the fourth section details selected tour formation algorithms and the experimental setting; the fifth section presents the results of the experiments to compare the algorithms, as well as the prediction of day-to-day pattern homogeneity and the differences across sub-populations; the sixth section concludes this paper, summarising obtained insights for data processing and modelling practice.

Literature Review

We define some terminology for purposes of this research. Vehicle *trip ends,* also known as vehicle *stops*, can be grouped into those *relevant* and those *nonrelevant* to the analysis at hand. For example, stops for short breaks might be considered differently than those to deliver goods. Zhou et al. (2014) justify this classification method. The sequence of trips taken between two *relevant stops* are defined as *trip chains* (Holguín-Veras and Patil 2005), represented in Fig. 1 by ({a}, {b, c}, {d}, {e}, {f}). A *tour* consists of one or more *trip chains*. For illustrative purposes, in Fig. 1, assume a return to *relevant stop* numbered as "1" marks the end of a *tour*. Then, there are two *tours* in Fig. 1: the first composed by *trip chains* ({a}, {b, c}) and the second by *trip chains* ({d}, {e}, {f}). A *tour-chain* is defined as the set of *tours* that occurs within a day. We borrow this term from Ruan et al. (2012). In Fig. 1, the *tour-chain* is composed of two *tours*.

With regard to data processing, the critical first steps rely on the methods for identifying stops from trips and their

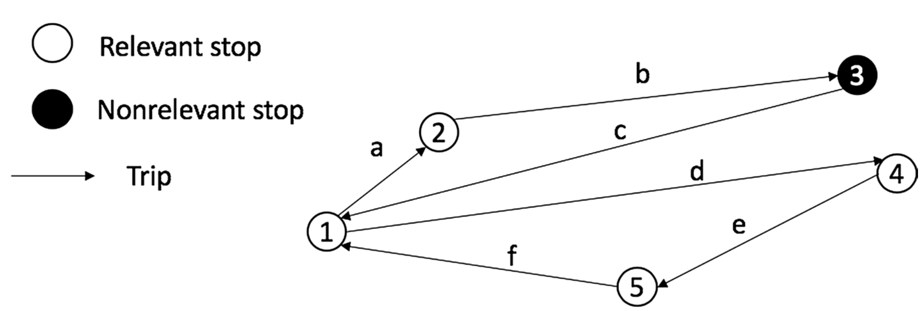

Fig. 1 Example sequence of stops and trips

purposes (Du and Aultman-Hall 2007; Greaves and Figliozzi 2008; Schuessler and Axhausen 2009, Sharman and Roorda 2011; Joubert and Axhausen 2011, Yang et al. 2014). Data might be collected to inform the activities at/and destinations (Alho et al. 2018). Even in the case such data is not available (i.e. only GPS traces are available), activities and destinations can be inferred. Examples are given by Sharman et al. (2012) who identify depots from the attributes of stops, and Sharman and Roorda (2013) who propose a process to assign parcel level information about the destinations.

Following the identification of stops, there are two seminal quantitative studies that address stop-to-tour assignment. Liedtke and Schepperle (2004) briefly describe a fuzzy logic-based pattern recognition method to process a 1.7 million trip records in five tour-types, covering both urban and interurban trips. You and Ritchie (2018) propose a method to post-process GPS data for identifying freight vehicle tours and applied it to GPS traces of freight vehicles traveling from/to port facilities. Beziat et al. (2015) presents tour-type identification by qualitatively defining 13 tour profiles based on interviews and a review of academic literature.

Several studies focus on the relationships between tour-type and tour-chain and driver, vehicle, shipment, and operator characteristics. Zhou et al. (2014) assume tour-types as per the total of deliveries performed in a day and show that commercial vehicle tour-types tend to be associated with commodity type, land use type, loading/unloading cargo weight and travel speed. Ruan et al. (2012) identify five major daily tour-chain-types based on the vehicle base location(s), tours per day and stops per tour. Using urban commercial vehicle survey data for tour-chaining choice model estimation, they analyse the relationship between tour-chain-types, cost and shipment characteristics. Khan and Machemehl (2017) also determine tour-chain-types as function of a base location, tours per day and stops per tour. They estimate a model for determining tour-chain-types and the number of trips, based on a wide list of factors using a multiple discrete–continuous extreme value model. The aforementioned studies do not reveal the details of the processes that lead to defining the tour-chain-types. Sharman and Roorda (2011) analyse day-to-day variations in terms of the overlap between the stop locations. While their data indicates that few destinations were visited on a daily basis, the research does not cover the analysis into the regularity of tour-type or tour-chain-type in freight vehicle operations; such analysis could provide insights on vehicle operational homogeneity across vehicle type or industry type over a certain period, informing whether single-day sampling would be sufficient for obtaining data about the routine of freight vehicle operations. As for the usage of the processed data, You et al. (2016) present a modelling framework of freight flows with spatial–temporal constraints that relies on tour-level data for calibration. Subsequently, You and Ritchie (2018) use tour-level data to explore tour-level behaviour of clean drayage trucks, revealing distinct travel patterns across days despite tour-types having repetitive patterns.

This review demonstrates the wealth of research that both contributes to and leverages tour-level analysis. In light of some gaps, we argue positively for a comparative study of methods applicable to stop-to-tour assignment, as well as of tour-type and tour-chain identification. In the present research, we aim to reveal (1) insights into the interpretation and inference of base locations; (2) the outcome of different assumptions in the algorithms to identify freight vehicle tours from stop chains; (3) the outcome of different assumptions in the algorithms to identify tour-type and tour-chain-types; and (4) day-to-day pattern homogeneity with regard to tour-type and tour-chain-type for a sample of tracked vehicles.

Data

The data used in this research consists of GPS traces and stop-level data obtained from a driver survey. The data collection process and the data collection platform, Future Mobility Sensing, are extensively described in Alho et al. (2018) and You et al. (2018), respectively. Stop-level data include stop purpose, location, duration, and cargo volume handled. The dataset includes records of 2151 driver-days with 497 unique drivers/vehicles. The sample is summarized by the vehicle body type and by industry type served in Tables 1 and 2, respectively. It should be noted that the

Table 1 Samples by vehicle body type

Body type	Sample size	Share (%)	Extraction rate (%)
Low loader	1	< 1	< 1
Recovery vehicle	2	< 1	< 1
Refrigerated vehicle	3	< 1	3
Concrete/cement mixer	2	< 1	1
Platform truck	11	2.21	< 1
Tanker	7	1.41	< 1
Crane	2	< 1	< 1
Lorry metal	9	1.81	24
Garbage/sanitary wagon	22	4.43	1
Van	15	3.02	36
Prime mover	90	18.11	3
Wooden frame lorry	90	18.11	6
Tipper/dump truck	204	41.05	3
Unknown[a]	39	7.85	–

–, not applicable

[a]Body types of some samples are unknown due to incomplete survey responses

Table 2 Sample by vehicle operations industry

Industry	Sample size[a]	Share (%)
Accommodation	5	1
Agriculture	1	< 1
Construction	354	58
Manufacturing	60	10
Mining	1	< 1
Other services	10	2
Retail—food and beverage	5	1
Retail—non food and beverage	14	2
Transportation and storage	93	15
Utilities and waste	21	3
Wholesale	7	1
Unknown[b]	43	7

[a]The sum of the values is above 497 as the same respondent might serve multiple industries

[b]Due to incomplete survey data

sample is not representative of the population of freight vehicles in Singapore and was not collected with such intention. The data is only used to showcase an application of the methods further described. Although the above-mentioned data is not big data in itself, the algorithms could be applied to big data and the insights we intend to provide aim to inform a purposeful application. Stop-level data are often unavailable for GPS traces (Holguín-Veras and Patil 2005; Eluru et al. 2018) as driver surveys for stop-level data are costly. However, stop-level data are inferable (Sharman et al. 2012). Furthermore, for data collected by fleet monitoring system, stop-purpose inference algorithms can leverage small surveys and/or Point of Interest data. Moreover, if GPS traces are collected by vehicle operators, these data could be matched with the activity and destination information based on their shipment records.

Methods

Stop Identification

The identification of *stops* is a two-step process using a custom developed method (Zhao et al. 2015) which includes DBSCAN (Ester et al. 1996), a clustering algorithm. First, a stop detection algorithm is applied, and then we aggregate raw stop records over the vehicle tracking period. Specifically, at the vehicle level, stop records at *nearby* coordinates, within 500 m, are considered the same stop. As mentioned earlier, we assume as *relevant stops* those for deliveries and/or pickups as well as those at a *base*. Any other stops are considered *nonrelevant*.

Base Identification

The definition of "base" can vary across freight agents (such as shippers or carriers). This understanding seems to be shared by Ruan et al. (2012) who hypothesize multiple functions for the base, such as a "distribution center, a warehouse, a business location (e.g., retail store, construction site), or fleet operator's home office/garage". Furthermore, for a given driver/vehicle, there could exist multiple bases which differ not only in purpose but also in location. For example, a driver/vehicle might have a "parking base", i.e. overnight parking location, and a "pickup base", i.e. the facility to which the vehicle returns multiple times during the day for picking up the goods. Ruan et al. (2012) also propose tour-chain structures that consider multiple bases. Selecting either of these bases as the pivotal point of tours potentially leads to a different set of tours despite the same stops being visited in the same order. We aim to contribute to this identification process, by exploring other algorithms that consider the purpose of stops and/or vehicle payload, as it will be further explained.

Base identification was prior addressed by Sharman and Roorda (2011) in the context of having no data apart from raw GPS traces. The authors evaluated the existence of bases (named 'depots') by considering several variables. Selection criteria were related to the percentage of stops performed at a given location within the study area and the average duration of the longest stop at such location on sampled days. Although the method is applicable to our case, our process differs as we attempt to leverage survey data first, providing an illustration of an alternative process.

In the driver survey we leverage, drivers had the option to declare a *frequent stop* (location) as a *base*, subject to their perception of what a *base* is. In our method, we first attempt to leverage declared bases over those identified using other methods, if confirmed "true" according to the criteria of "daily visits". If there is a need for base identification, we first aim to use the locations, where drivers change shifts daily, and only subsequently, locations, where pickups occur daily. The main justification for the latter is that one of the stop-to-tour assignment methods uses pickup locations, and we wish to keep the applications distinct. The algorithm is described by the following high-level pseudocode, and further detailed in "Appendix 1":

```
For each driver:
      Iterate through the respondent's declared frequent stops:
            If a stop is a declared base AND has been visited on all days of
            the verified detected stops:
                  Add the stop to base list

      If base list has no entries:
            Iterate through the respondent's detected stops:
                  If the purpose at the stop is "Change Shift" AND has been
                  visited on all verified days:
                        Add the stop to base list

      If base list has no entries:
            Iterate through the respondent's detected stops:
                  If the purpose at the stop is "Pickup" AND has been visited
                  on all verified days:
                        Add the Stop to base list
```

Stop-to-Tour Assignment

We explore three algorithms for tour-type identification purposes, focusing on the regularity of activities, the type of activities performed, and the vehicle capacity usage. These are not an exhaustive list and other variables could be explored, which is out of the scope of this research, as their applicability is not so clear. All algorithms iterate over the stop sequences, inspecting the characteristics of each stop and assigning a sequential tour-identification number to it. These are:

- *Base-driven* algorithm: A return to the identified base marks the end of a tour. This algorithm is most aligned with prior research applications such as You and Ritchie (2018) and Gopalakrishnan et al. (2019).
- *Purpose-driven* algorithm: A pickup stop that follows a delivery marks the start of a new tour. This algorithm aligns to the case, where a vehicle returns, or heads to, one or more "operational" base(s) for picking up goods throughout the day. It has been applied by Jing et al. (2019) and Alho et al. (2019).
- *Capacity-driven* algorithm: A pickup stop by an empty vehicle (i.e. the capacity usage is zero or equal to zero) marks the start of a new tour. We expect some level of alignment between the outputs from *Capacity-driven* and *Purpose-driven* algorithm, unless vehicles do pickups with some of the prior load still in the vehicle.

In all the three algorithms, a stop that follows a prior stop with a duration of over 240 min is considered the start of a new tour. This threshold was defined similarly to past research (You and Ritchie 2018) and observed in our data as a clear point demarcating between stop durations during operation periods (e.g., those for rest, pickup, and delivery) and those during non-operation periods (e.g., overnight and over the weekend). The high-level pseudocode for the algorithms is described following and further detailed in "Appendix 2".

The *Base-driven* algorithm pseudocode is:

```
Tour id = 0
Iterate over stop sequence
      If stop location is in the list of bases OR (prior stop duration > 240
      minutes)
            Increase tour id by 1
      Else
            Assign prior tour id
```

The *Purpose-driven* algorithm pseudocode is:

Table 3 Illustration of hypothetical application of stop-to-tour assignment algorithms

Stop id	1	2	3	4	5	7	8
Base	Yes	No	No	No	Yes	No	No
Purpose	P	P	D	D	P	D	P
Capacity usage (%)	50	100	50	25	75	0	100
Tour id (base-driven algorithm)	1	1	1	1	2	2	2
Tour id (purpose-driven algorithm)	1	1	1	1	2	2	3
Tour id (capacity-driven algorithm)	1	1	1	1	1	1	2

```
Tour id = 0
Iterate over stop sequence
      If (current stop purpose includes Pickup AND previous stop purpose
      includes Delivery) OR (prior stop duration > 240 minutes)
            Increase tour id by 1
      Else
            Assign prior tour id
```

The *Capacity-driven* algorithm pseudocode is:

```
Tour id = 0
Iterate over stop sequence
      If (net vehicle capacity usage = 0 and stop purpose includes Pickup) OR
      (prior stop duration > 240 minutes)
            Increase tour id by 1
      Else
            Assign prior tour id
```

Table 4 Tour-type identification criteria

Tour-type	No. of pick-ups/tour	No. of pickup loc.	No. of deliv-eries/tour	No. of delivery loc.
Direct	1	1	1	1
Unloading	1	1	> 1	> 1
Loading	> 1	> 1	1	1
Mixed	> 1	> 1	> 1	> 1

Table 3 illustrates the algorithms' application to a hypothetical case. It can be seen that for the *Base-driven* algorithm tour-identification number (id) switches from 1 to 2 upon visiting the base. In this case, the increment is aligned to the non-sequential pickup in the *Purpose-driven* algorithm, whereas the *Capacity-driven* algorithm only triggers a tour id change when the vehicle capacity reaches zero (stop 8).

We compare the outputs of the three stop-to-tour algorithms in terms of the mean and standard deviation (SD)

Table 5 Tour-chain identification criteria

Tour-chain group	Tour-chain	No. of tours	No. of pickups/tour	No. of pickup loc.	No. of deliveries/tour	No. of delivery loc.
Direct	Single direct	1	1	1	1	1
	Fixed pickup, fixed delivery	> 1	1	1	1	1
	Fixed pickup, unfixed delivery	> 1	1	1	1	> 1
	Unfixed pickup, fixed delivery	> 1	1	> 1	1	1
	Unfixed pickup, unfixed delivery	> 1	1	> 1	1	> 1
Unloading	Single unloading	1	1	1	> 1	> 1
	Fixed pickup, multiple fixed deliveries	> 1	1	1	> 1	> 1 and fixed[a]
	Fixed pickup, multiple unfixed deliveries	> 1	1	1	> 1	> 1
	Unfixed pickup, multiple fixed deliveries	> 1	1	> 1	> 1	> 1 and fixed[a]
	Unfixed pickup, multiple unfixed deliveries	> 1	1	> 1	> 1	> 1
Loading	Single loading	1	> 1	> 1	1	1
	Multiple fixed pickups, fixed delivery	> 1	> 1	> 1 and fixed[a]	1	1
	Multiple unfixed pickups, fixed delivery	> 1	> 1	> 1	1	1
	Multiple unfixed pickups, unfixed delivery	> 1	> 1	> 1	1	> 1
Mixed	Single mixed	1	> 1	> 1	> 1	> 1
	Multiple mixed	> 1	> 1	> 1	> 1	> 1

[a]Fixed means location set (i.e., the locations of one or more stops) remains the same across tours; *unfixed* means location set varies across tours

of the following indicators: (1) tours per day, (2) stops per tour, (3) tour duration (minutes), and (4) tour distance (kilometres). Our intention is not to reveal *the best* algorithm but rather to expose the implications of selecting them. To the best of our knowledge, there is no consensus on how to evaluate the effectiveness of the algorithms in revealing the "true" tours, since the concept is a human construct. In fact, as mentioned earlier, different interpretations of tours are used for different applications.

Tour-Type and Tour-Chain Identification

In the past research, tour-type and tour-chains have been defined considering base location, tours per day and stops per tour (Ruan et al. 2012, Khan and Machemehl 2017). This is considered as a valid method but potentially relies on a subset of relevant variables. Therefore, we explore also stop purposes and their recurrence, as well as regularity of stop locations. Regarding stop-purpose recurrence, the distinction is made in terms of the number of pickups relative to the number of deliveries. It is based on the hypothesis that certain combinations of stop-purpose recurrence are strongly correlated with operational characteristics of freight movements. Regarding the regularity of stop locations, we make a distinction between those stops locations that are *fixed* and *unfixed*. *Fixed* stop locations are those visited more than once in a day, while those *unfixed* are only visited once in a day.

The definitions of tour-types and tour-chains to be explored are shown in Tables 4 and 5, respectively. The identification process primarily categorizes tours and tour-chains into the following four groups:

- *Direct* tours that consist of one pickup and one delivery, associated with full truck load (FTL) shipments. Past research indicates that this tour-type is associated with longer distances travelled and larger dwell times (Ruan et al. 2012).
- *Unloading* tours that consist of one pickup and more than one delivery, associated with less than truckload shipping (LTL) (e.g., parcel deliveries).
- *Loading* tours that consist of more than one pickup and one delivery, associated with operations such as those of waste collection.
- *Mixed* tours that consist of multiple pickups and deliveries, and can be associated with delivery tours that also collect returned shipments.

To identify the tour-chain, two algorithms are considered:

- *Tour-type-based identification (TT)* Tour-chain is identified based on the types of tours performed in a day. If a tour-type accounts for at least 60% of all tours within a day, the tour-chain is labelled by such tour-type. 60% is set assuming that when two tours of different types are performed daily there is no predominant type.
- *Tour-chain-based identification (TC)* This alternative algorithm characterizes the tour-chain at the day level. Instead of using the predominant tour-type, the algorithm reads stops-to-tour assignments, averages the stops per tour by purpose and then identifies tour-chain for a day.

Note that tour-chain groups are also defined, consistent with the ratios between #Pickups/tour and #Deliveries/tour, for achieving direct comparisons between algorithms and further use in day-to-day pattern homogeneity analysis.

Day-to-Day Pattern Homogeneity Analysis

The day-to-day pattern homogeneity analysis is used as a partial demonstration of how differences in the assumptions can lead to differences in outputs. Moreover, it provides insights on whether there is some level of homogeneity on the patterns performed by the vehicles.

We propose to use an entropy concept (Eq. 1) to quantify day-to-day pattern homogeneity. In past research, the entropy concept was applied to calculate the diversity of commercial establishment functions (Alho and de Abreu e Silva 2014). In this research, we apply it to measure the diversity of tour patterns. Pj is the proportion of tours (or tour-chain groups) of type j. J is defined in this case as the number of tour-types or tour-chain groups, depending on the application. This indicator is normalized and, therefore, ranges between one (with the equal share across tour-types

Table 6 Tasks and run time

Task	Run time
Clustering (DBSCAN)	~4 min 51 s
Purpose/capacity-driven tours identification and classification (TT)	~1.5 s
Purpose/capacity-driven tours identification and classification (TC)	~2 s
Base-driven tours identification and classification (TT)	~34 s
Base-driven tours identification and classification (TC)	~34 s
Day-to-day pattern homogeneity calculation	~1 s

or tour-chain groups) and zero (with the presence of only one tour-type or tour-chain groups):

$$\text{Entropy} = \sum_{j} \frac{\left|P_j \times \ln\left(P_j\right)\right|}{\ln(J)}. \tag{1}$$

The entropy is measured for the results of both algorithm outputs in tour-chain-type identification (TT and TC). It should be noted that the method is applied differently in both cases. For TT, we group tours by tour-types across the observed period, since the generalization at the daily level (i.e. the definition of the tour-chain) lowers the resolution of the inputs. For TC, we simply use tour-chain-type group. Thus, TT is applied from a perspective of all tours over the period (i.e., one or more tours types per day), while TC inputs are at the daily level (i.e., one tour-chain-type group per day).

Software and Hardware

To process the data, we use scripts written in the Python programming language. The selected hardware was an Intel Core i7-7700 CPU @ 3.60 GHz processor and 32 Gb RAM. In the current experimental setting, the algorithms run under a batch processing model, with run times, as listed in Table 6. The algorithms are compatible with the latest developments in the FMS platform (You et al. 2018), which has been developed to process and display collected GPS data from several types of loggers (stand-alone, built-in smartphones, and tablets). The scripts can be coupled in the platform as a streaming model.

Results

Base Identification

As mentioned earlier, we start with a comparison between bases declared by the respondent and those detected from the GPS traces by the detailed algorithm. This process allows us to clarify whether there is some common understanding of "base locations" among drivers and whether revealed data is valuable when compared with inferred data.

500 drivers declared 1072 *frequent stops*, out of which 502 *frequent stops* were marked as *bases,* an average of 1 *base* per respondent. 66% of declared *bases* were the locations, where drivers start/end the work shift, followed by the locations, where cargo is picked up (16%). On the other hand, 63% of the non-base *frequent places* were associated with locations, where cargo and/or trailers are picked up. This revealed that some level common understanding of the concept of "base" exists, with it being, where the work shift starts/ends, and not necessarily, where regular pickups are performed. Despite this, and contrary to expectations, *frequent stops* were found to be visited sparingly. During the period of the survey (5 days), 35% of *declared base frequent places* and 6% of *declared non-base frequent places* were visited, which highlights some fundamental flaw either in the process of recalling or reporting information. This result indicates that declared bases (or frequent places) are not suitable as the reference points to identify tours.

Following these conclusions, we set to use revealed locations which were often visited as bases. Out of 8718 detected stops (i.e. clusters of raw stop records), 1186 were visited every day. Out of these stops with non-mutually exclusive activities, 88% are associated with a start/end work shift activity, 42% a pickup and 40% a delivery. Following, for 93% of drivers a single base was identified, and 2% of drivers have two or more bases identified. The records for which no bases were identified (5% of drivers), were excluded from the following steps of the analysis. For those drivers with base(s) identified, 34% were from declared bases, 57% were from revealed locations, where drivers change shift frequently, and 9% were from revealed locations, where drivers perform pickups frequently.

Stop-to-Tour Assignment

Table 7 shows the results across the different stop-to-tour assignment algorithms. Relatively similar results can be observed in the *Purpose-driven* and *Capacity-driven* algorithms, whereas the *Base-driven* algorithm leads to different outcomes. As expected, the latter leads to a smaller average number of tours, with more stops per tour, since the bases are, in many cases, the locations, where the drivers start/end the work shift (Fig. 2). The alignment between the *Purpose-driven* and *Capacity-driven* algorithms is mainly due to that

Table 7 Comparison of stop-to-tour assignment algorithm results

Stop-to-tour assignment algorithm	Tours per day		Stops per tour		Tour duration (min)		Tour distance (km)	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Base-driven	1.39	1.11	8.51	6.12	211.56	152.87	121.28	95.03
Purpose-driven	3.88	2.47	3.03	1.66	74.51	72.45	42.89	30.74
Capacity usage-driven	3.42	2.41	3.44	2.79	84.45	88.89	48.81	42.21

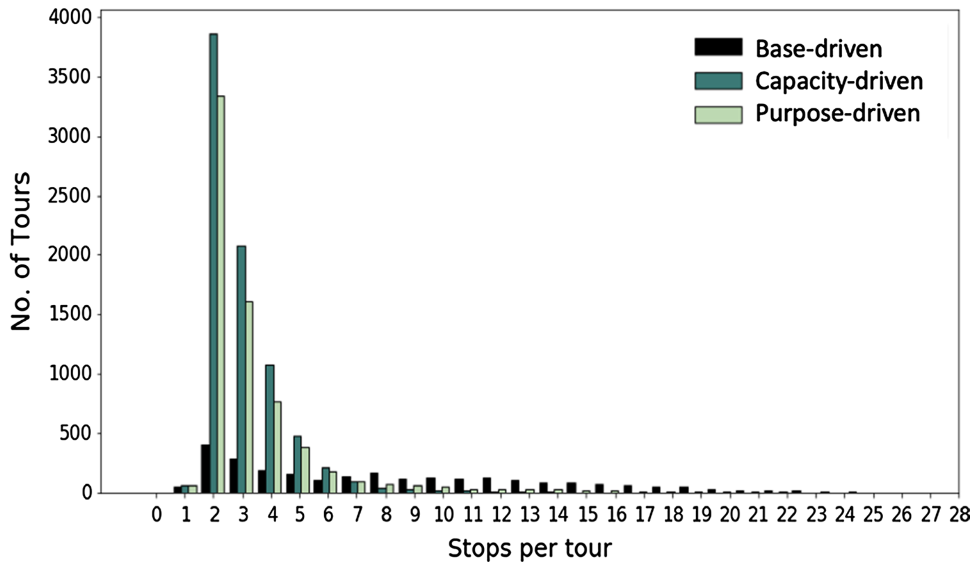

Fig. 2 Frequency of stops per tour for each algorithm

Table 8 Tour-chain identification results using TT

Tour-type/tour-chain-type	Base-driven		Purpose-driven		Capacity-driven	
	Tours (%)	Tour-chains (%)	Tours (%)	Tour-chains (%)	Tours (%)	Tour-chains (%)
Direct	23.6	10.9	81.0	72.6	80.2	63.5
Unloading	4.3	2.2	5.7	3.3	4.4	2.7
Loading	3.7	1.7	4.0	1.4	1.8	0.9
Mixed (Mixed Single/Multiple)	57.2	62.6	5.5	4.5	10.6	13.1
Non-identifiable	11.3	22.5	3.8	18.2	3.0	19.8

Table 9 Tour-chain identification results using TC

Group	Tour-chain-type	% Tour-chain		
		Base-driven	Purpose-driven	Capacity-driven
Direct	Direct (single tour)	7.6	7.8	7.4
	Fixed pickup, fixed delivery	1.1	10.8	9.8
	Fixed pickup, unfixed delivery	0.9	12.7	10.2
	Unfixed pickup, fixed delivery	0.2	4.3	3.7
	Unfixed pickup, unfixed delivery	2.0	38.8	30.6
	Total	11.8	74.4	61.6
Unloading	Unloading (single tour)	1.5	1.6	1.6
	Fixed pickup, multiple fixed deliveries	0.1	0.0	0.0
	Fixed pickup, multiple unfixed deliveries	0.6	1.5	0.9
	Unfixed pickup, multiple fixed deliveries	0.0	0.1	0.0
	Unfixed pickup, multiple unfixed deliveries	0.3	2.5	1.4
	Total	2.6	5.8	4.0
Loading	Loading (single tour)	0.9	1.1	1.1
	Multiple fixed pickups, fixed delivery	0.1	0.2	0.0
	Multiple unfixed pickups, fixed delivery	0.1	0.2	0.0
	Multiple unfixed pickups, unfixed delivery	0.1	1.4	0.7
	Total	1.2	3.0	1.8
Mixed	Mixed (single/multiple)	46.5	7.5	15.8
Non-identifiable		38.0	9.2	16.6

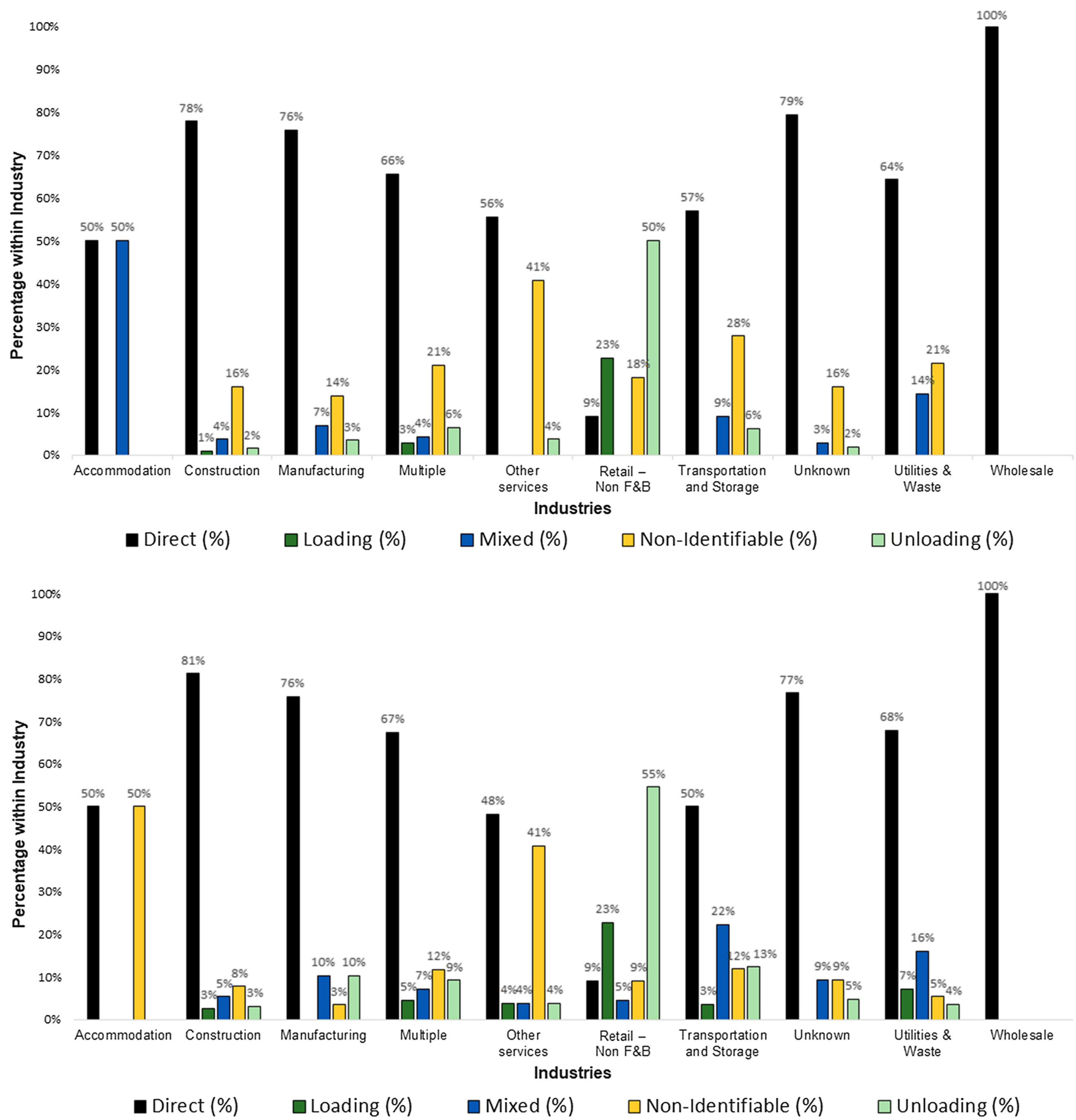

Fig. 3 Tour-chain-type groups and industry served by vehicle (top: TT; bottom: TC)

fact that most vehicles return/head to the next pickup location with an empty load (i.e., perform full truck load operations). The choice of the algorithm has an influence on the tour-level indicators, and the base-driven algorithm is prone to reveal longer tours potentially including loading/unloading operations within the tour-chain.

Tour-Type and Tour-Chain Identifications

Tables 8 and 9 show the results for tour-type and tour-chain identifications, respectively, using TT and TC. In Table 8, the results in the *Tours* column shows the share of each tour-type identified, whereas the *Tour-chain* column illustrates the outcome of the TT process using the 60% threshold.

The results of the TT application provide interesting insights. The *Base-driven* algorithm, associated with predominantly change-shift locations leads to a much higher share of Mixed tours, i.e. containing multiple pickups and deliveries, than the *Purpose-driven* and *Capacity-driven* algorithm. The tours identified by *Purpose-driven* algorithm are more compatible with the tour-type alternatives, since it relies on the stop activity purpose. The *Capacity-driven* algorithm was expected to produce similar results to *Purpose-driven* algorithm, as prior data analysis revealed that vehicles load fully at the pickup stops, and this seems to hold in many cases. For these algorithms, it can be noticed that there is a larger share of Non-identifiable tour-chains versus non-identifiable tours. This is due to situations, where there are no predominant tour-types, particularly associated with large share of days illustrated as having two tours in Fig. 2.

Regarding TC algorithm, the results follow a similar pattern, particularly in the alignment between the results for the *Purpose-driven* and *Capacity-driven* algorithm. As expected, the share of Mixed and Non-identifiable tours reveals that the *Base-driven* algorithm does not allow

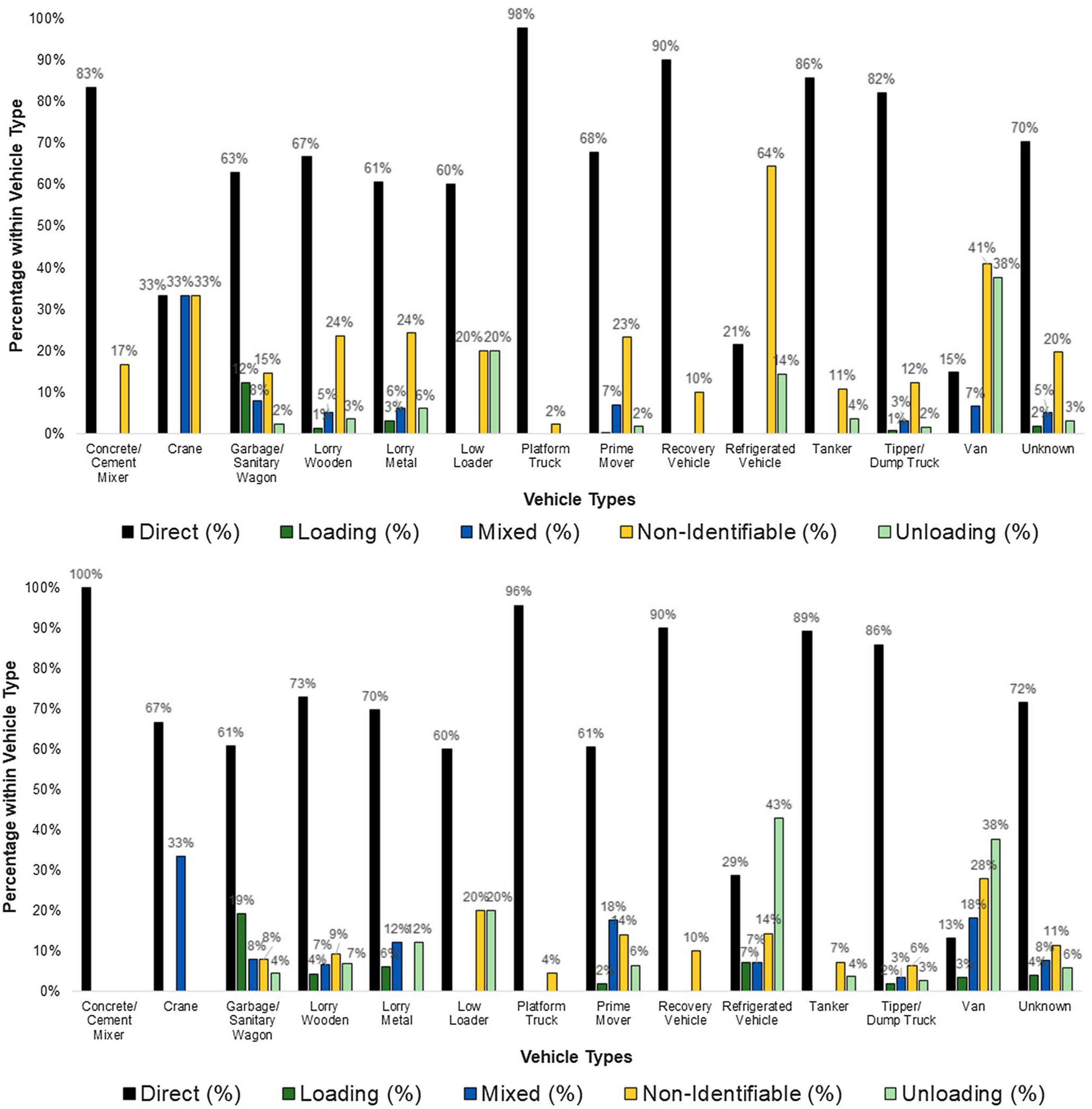

Fig. 4 Tour-chain-type groups and vehicle body type (top: TT; bottom: TC)

understanding fully the patterns of pickups and deliveries performed by the vehicles.

There are interesting algorithm-to-algorithm comparisons that can be drawn and are demonstrated for an application of the *Purpose-driven* method. Regardless of the application of the TT or TC algorithm, most of tour-chains belong to the same group (direct, unloading, loading, or mixed). Specifically, matches for direct group are 94%, for Unloading 86%, for Loading 86%, and for Mixed 89%. Differences are found in *Non-identifiable* tours, with 37% matched across algorithms.

An advantage of the TC algorithm is the ability to reveal that direct tours–chains have different natures, even for relatively homogeneous samples like the one used. For example, considering tour-chains in the direct group, and using the purpose-driven algorithm, 39% of the cases are "Unfixed pickup, unfixed delivery". This reveals a non-negligible share of tours that are not what direct tours are intuitively associated with; it is often thought that, in many cases, a fixed distribution center is used for full truck load shipments to several destinations. Another interesting finding is that the TC algorithm produces results that reveal: (a) less predominance of direct tours and (b) smaller share of Non-identifiable tour-chains. This was expected, since averaging out stops per tour "smoothens" tour heterogeneity at the daily level. The application of TC allows for a decrease in approximately 63% of tours–chains that the TT could not identify. Lacking "ground truth", no process can ultimately be deemed as correct or wrong. However, deeming higher output resolution as desirable, the TC algorithm seems to be more suitable.

Following, we detail the results from the perspective of industry served and vehicle body type, selecting the *Purpose-driven* algorithm due to its better fit to the TC/TT methods as well as high replicability potential compared with the *Capacity-driven* algorithm. We aggregated the outputs of the TC algorithm using the prior defined groups to allow for a comparison with the TT outputs. Similar proportions of tour-chains, at the group level, can be observed for sampled vehicle-associated industry types (Fig. 3) with most industries operating on direct tours. The retail industry stands out (albeit the small sample size), as over 70% of its tour-chains are Loading and Unloading tours. Regarding the tour-types associated with the sampled vehicle types (Fig. 4), direct tours–chains were also predominantly observed in most cases, other than Refrigerated Vehicles and Vans. This is not surprising, since about half of the vans, and one-third of the refrigerated vehicles serve the retail industry.

Day-to-Day Pattern Homogeneity Analysis

In the application of the tour identification with the *Purpose-driven* method, we compare the day-to-day pattern homogeneity of the outputs from algorithm TT and TC, disaggregated by associated industry type (Table 10) and vehicle body type (Table 11). Industry and vehicle body types that show high (or low) entropy values are consistent between the two algorithms but entropy values tend to

Table 10 Entropy of tours quantified across industries served by vehicles

Industry	Sample size	TT		TC	
		Mean	SD	Mean	SD
Accommodation	1	0.50	–	0.33	–
Construction	296	0.23	0.18	0.14	0.18
Manufacturing	13	0.27	0.14	0.20	0.18
Retail—non F&B	5	0.26	0.29	0.19	0.29
Transportation and storage	36	0.29	0.21	0.21	0.19
Utilities and waste	13	0.33	0.13	0.23	0.24
Wholesale	1	0.31	–	0.00	–
Other services	7	0.30	0.18	0.21	0.21
Multiple	94	0.23	0.19	0.20	0.21
Non-identifiable	25	0.25	0.12	0.16	0.20

–, non-applicable

Table 11 Entropy of tours quantified across vehicle body types

Vehicle body type	Sample size	TT		TC	
		Mean	SD	Mean	SD
Refrigerated vehicle	3	0.51	0.08	0.38	0.34
Low loader	1	0.38	–	0.46	–
Van	15	0.35	0.27	0.21	0.23
Garbage/sanitary wagon	22	0.26	0.19	0.16	0.21
Prime mover	87	0.27	0.18	0.21	0.20
Lorry wooden	90	0.26	0.20	0.21	0.22
Lorry metal	9	0.23	0.20	0.15	0.15
Recovery vehicle	2	0.22	0.02	0.12	0.17
Concrete/cement mixer	2	0.24	0.13	0.00	0.00
Tipper/dump truck	202	0.21	0.15	0.11	0.16
Tanker	6	0.15	0.13	0.12	0.19
Crane	2	0.15	0.22	0.00	0.00
Platform truck	11	0.12	0.15	0.03	0.09
Non-identifiable	39	0.26	0.20	0.22	0.21

–, non-applicable

be higher for the application of TT compared with that of the TC.

Looking at the tour-type homogeneity, in terms of industries served, construction is an example of having more homogeneous tour-types. With regard to vehicle body types, vehicles predominantly associated with construction also demonstrate this behaviour, e.g., Tipper/Dump Trucks.

Conclusions

A research gap has been identified regarding the methods of processing freight vehicles' GPS data to tour and tour-chain data. The main steps to which this research aims to contribute are stop-to-tour assignment as well as tour-type and tour-chain-type identifications. In this paper, we explored several algorithms for such objectives, i.e., assigning stops to tours, identifying tour-types and tour-chain-types, and compared their outputs, highlighting differences. A major finding about the identification of bases, a critical input to one of the stop-to-tour assignment algorithms (*Base-driven*), was that declared data regarding bases might not be as accurate as inferred data. This finding holds despite a common understanding that the base is mostly associated with the place, where drivers change shift, which are not necessarily pickup locations. This also contributes positively to arguments towards not fully relying on survey data, which is likely to occur for applications to big data. Our analysis also revealed that most vehicle operations were associated with a base visited daily, in line with the most common assumptions in the literature regarding tour starting/end points. However, these findings also impact on the application of the subsequent tour-type and tour-chain-type identification algorithms. The *Base-driven* and *Purpose-driven* algorithms relied on different types of pivotal points to trigger the start of a new tour, resulting in considerably different outputs (tour counts and number of stops per tour). Such difference resulted in a low compatibility of the identified tours using the *Base-driven* algorithm with the selected tour-type and tour-chain-type identification methods. Despite this, we highlight that this conclusion could be related to the nature of the data used in the application, and further applications are recommended. In case an alignment between identified bases and pickup locations had been achieved, the results would be expectedly different. Thorough this paper, several differences in outputs arising from a combination of the methods selected and data at hand have been exposed. Ultimately, our findings indicate that researchers should take due diligence on selecting algorithms and provide clear descriptions on the selected pre-processing steps for a better understanding from the readers on the potential implications of the assumptions.

Acknowledgements This research is partially supported by the Singapore Ministry of National Development and the National Research Foundation, Prime Minister's Office under the Land and Liveability National Innovation Challenge (L2 NIC) Research Programme (L2 NIC Award No. L2 NICTDF1-2016-1). We would also like to thank partner agencies, the Urban Redevelopment Authority of Singapore, the Land Transport Authority of Singapore and JTC Corporation for their support. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the Singapore Ministry of National Development, National Research Foundation, Prime Minister's Office, Singapore, the Urban Redevelopment Authority of Singapore, JTC Corporation, Land Transport Authority of Singapore.

Author contributions The authors confirm contributions to this paper as follows. Study conception and design: ARA, TS, KJ, PJ, and MBA; data preparation: ARA and MHC; analysis and interpretation of results: ARA, MHC, and TS; draft manuscript preparation: ARA and TS. All authors reviewed the results and approved the final version of the manuscript.

Appendix 1: Base Identification Algorithm

```
Inputs:
freqStopDeclared: A list of frequent stops with activity and location
declared by drivers.

baseDeclared: A list of boolean values (True/False) stating if each
stop location in freqStopDeclared is declared as base.

visitedAllDaysDeclared: A list of boolean values stating if each stop
location in freqStopDeclared has been visited at least once a day
throughout the survey period.

freqStopDetected: A list of frequent stop locations, detected from GPS
traces.

freqStopPurpose: A list of stop purposes corresponding to each stop
location in freqStopDetected.

visitedAllDaysDetected: A list of boolean values (True/False) stating
if each stop location in freqStopDetected has been visited at least
once a day throughout the survey period.

Output:
identifiedBase: A list of stop locations identified as base.


identifiedBase = [] #Initialise empty list
#Step 1:
For each location in freqStopDeclared
      If (corresponding value in baseDeclared is True) AND
      (corresponding value in visitedAllDaysDeclared is True)
            THEN add location to identifiedBase

#Step 2:
If identifiedBase is empty
      THEN For each location in freqStopDetected
                  If (corresponding purposes in stopPurpose contains
                  "Change Shift") AND (corresponding value in
                  visitedAllDaysDetected is True)
                        THEN add location to identifiedBase

#Step 3:
If identifiedBase is empty,
      THEN For each location in freqStopDetected}
                  If (corresponding purposes in stopPurpose contains
                  "Pickup") AND (corresponding value in
                  visitedAllDaysDetected is True)
                        THEN add location to identifiedBase
```

Appendix 2: Tour-Identification Algorithms

Inputs:
stopsList: A list of stops detected from GPS traces.

stopDurationList: A list of stop durations corresponding to stops in ***stopsList.***

stopPurposeList: A list of stop purposes corresponding to each stop location in ***stopsList***.

stopCapacityUsageList: A list of vehicle capacity usage corresponding to each stop. E.g., picking up or Delivering a full truckload of cargo at a particular stop will mean a vehicle capacity usage of respectively 1 or -1.

netCapacityUsageList: A list of cumulative vehicle capacity usage as of the corresponding stop. A pickup stop will increase the Net Vehicle Capacity Usage, whereas a delivery stop will decrease it by the capacity offloaded.

identifiedBase: A list of stop locations identified as base (the output from Base Identification process).

Output:
tourIdList: A tour Ids associated with each to stop location instance.

Base-driven algorithm:

```
#Initialise Tour Id at zero
tourId = 0

For each stop in stopsList
      If (corresponding stop location in stopsList is in
      identifiedBase) OR (prior stop's duration in stopDurationList >
      240 minutes)
            THEN Increase tourId by 1 #Stop belongs to a new tour.
                  Add tourId to tourIdList
      Else
            THEN Stop takes same tour Id as prior stop
                  Add tourId to tourIdList
```

Purpose-driven algorithm:

```
#Initialise Tour at zero
tourId = 0

For each stop in stopsList
      If (corresponding stop purpose in stopPurposeList includes
      "Pickup" AND prior stop's purpose includes "Delivery") OR (prior
      stop's duration in stopDurationList > 240 minutes)
            THEN Increase tourId by 1 #stop belongs to a new tour.
                  Add tourId to tourIdList
      Else
            THEN Stop takes same tour Id as prior stop
                  Add tourId to tourIdList
```

Capacity-driven algorithm:

```
#Initialise Tour at zero
tourId = 0

For each stop in stopsList
      If (corresponding net vehicle capacity usage in netCapacityUsageList
      is 0 AND prior stop's purpose in stopPurposeList includes
      "Delivery") OR (prior stop's duration in stopDurationList > 240
      minutes)
            THEN Increase tourId by 1 #stop belongs to a new tour.
                  Add tourId to tourIdList
      Else
            THEN Stop takes same tour Id as prior stop
                  Add tourId to tourIdList
```

References

Alho AR, de Abreu e Silva J (2014) Analyzing the relation between land-use/urban freight operations and the need for dedicated infrastructure/enforcement—application to the city of Lisbon. Res Transp Bus Manag 11:85–97. https://doi.org/10.1016/j.rtbm.2014.05.002

Alho AR, You L, Lu F, Cheah L, Zhao F, Ben-Akiva M (2018) Next-generation freight vehicle surveys: supplementing truck GPS tracking with a driver activity survey. In: 21st IEEE international conference on intelligent transportation systems, Maui, Hawaii

Alho A, Sakai T, Hara Y, Jing P, Raven M, Ben-Akiva M (2019) The influence of tour-formation assumptions on the reproducibility of freight vehicle flows. In: 7th International conference on city logistics, Dubrovnik, Croatia

Beziat A, Launay P, Toilier F (2015) A comprehensive view of goods transport systems—typology and analysis of freight tours in the Paris Region. In: 6th International urban freight conference, Long Beach, California

de Bok M, Tavasszy L (2018) An empirical agent-based simulation system for urban goods transport (MASS-GT). Procedia Comput Sci 130:126–133

Du J, Aultman-Hall L (2007) Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: automatic trip end identification issues. Transp Res Part A Policy Pract 41(3):220–232

Eluru N, Li X, Pinjari A, Abdel-Aty M, Anowar S, Momtaz SU, Iraganaboina NC, Keya N, Dey B, Zhao D, Balasu S, Sheela PV (2018) Freight data fusion from multiple data sources for freight planning applications. Final report, University of Central Florida, Department of Civil, Environmental and Construction Engineering, Orlando

Ester M, Kriegel H-P, Sander J, Xu X (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of 2nd international conference on knowledge discovery and data mining. AAAI Press, Portland, OR, pp 226–231

Gopalakrishnan R, Alho AR, Sakai T, Cheah L, Ben-Akiva M (2019) Assessing overnight parking infrastructure policies for commercial vehicles in cities using agent-based simulation. In: 7th International conference on city logistics, Dubrovnik, Croatia

Greaves S, Figliozzi M (2008) Collecting commercial vehicle tour data with passive global positioning system technology: issues and potential applications. Transp Res Rec J Transp Res Board 2049:158–166

Holguín-Veras J, Patil GR (2005) Observed trip chain behavior of commercial vehicles. Transp Res Rec J Transp Res Board 1906(1):74–80

Hunt JD, Stefan KJ (2007) Tour-based microsimulation of urban commercial movements. Transp Res Part B Methodol 41(9):981–1013

Jing P, Zhang Y, Jeong K, Alho AR, Ben-Akiva M (2019) Modeling daily tour-chaining pattern choice of urban heavy commercial vehicles. In: 98th Annual meeting of transportation research board, Washington, DC

Joubert JW, Axhausen KW (2011) Inferring commercial vehicle activities in Gauteng, South Africa. J Transp Geogr 19(1):115–124

Khan M, Machemehl R (2017) Analyzing tour chaining patterns of urban commercial vehicles. Transp Res Part A Policy Pract 102:84–97

Liedtke G, Schepperle H (2004) Segmentation of the transportation market with regard to activity-based freight transport modelling. Int J Logist Res Appl 7(3):199–218

Ruan M, Lin JJ, Kawamura K (2012) Modeling urban commercial vehicle daily tour chaining. Transp Res Part E Logist Transp Rev 48(6):1169–1184

Schuessler N, Axhausen KW (2009) Processing raw data from global positioning systems without additional information. Transp Res Rec J Transp Res Board 2105(1):28–36

Sharman BW, Roorda MJ (2011) Analysis of freight global positioning system data: clustering approach for identifying trip destinations. Transp Res Rec J Transp Res Board 2246(1):83–91

Sharman BW, Roorda MJ (2013) Multilevel modelling of commercial vehicle inter-arrival duration using GPS data. Transp Res Part E Logist Transp Rev 56:94–107

Sharman BW, Roorda MJ, Habib KMN (2012) Comparison of parametric and nonparametric hazard models for stop durations on urban tours with commercial vehicles. Transp Res Rec J Transp Res Board 2269(1):117–126

Short J (2014) Truck GPS data for tracking freight flows. In: Developing freight fluidity performance measures: supply chain perspective on freight system performance, pp 35–38. Transportation Research Circular Number E-C187. Transportation Research Board of the National Academies

Yang X, Sun Z, Ban XJ, Holguín-Veras J (2014) Urban freight delivery stop identification with GPS data. Transp Res Record J Transp Res Board 2411(1):55–61

You SI, Ritchie SG (2018) A GPS data processing framework for analysis of drayage truck tours. KSCE J Civ Eng 22(4):1454–1465

You SI, Chow JY, Ritchie SG (2016) Inverse vehicle routing for activity-based urban freight forecast modeling and city logistics. Transportmetr A Transp Sci 12(7):650–673

You L, Zhao F, Cheah L, Jeong K, Zegras C, Ben-Akiva M (2018). Future mobility sensing: An intelligent mobility data collection and visualization platform. In: 21st IEEE international conference on intelligent transportation systems, Maui, Hawaii

Zhao F, Ghorpade A, Pereira FC, Zegras C, Ben-Akiva M (2015) Stop detection in smartphone-based travel surveys. Transp Res Procedia 11:218–226

Zhou W, Chen Q, Lin JJ (2014) Empirical study of urban commercial vehicle tour patterns in Texas. In: 93th Annual meeting of transportation research board, Washington, DC