Transportation Research Part C 137 (2022) 103588

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Evaluation of methods to distinguish trips from activities in walking and cycling GPS data

Elmira Berjisian [a], Alexander Bigazzi [b,*]

[a] Department of Civil Engineering, University of British Columbia, Canada
[b] Department of Civil Engineering, University of British Columbia, 6250 Applied Science Lane, Vancouver BC V6T1Z4, Canada

ARTICLE INFO

Keywords:
GPS data
Cycling
Walking
Trip identification

ABSTRACT

An abundance of GPS data on walking and cycling requires substantial data processing – a primary component of which is trip identification to distinguish between data recorded during travel versus activities. The objective of this paper is to systematically evaluate trip identification algorithms in the literature and provide recommendations to improve performance for walking and cycling trips. Fourteen algorithms are applied to 1685 GPS trajectories from Vancouver, Canada, and evaluated on the bases of their agreement, distinction of trip and activity data characteristics, processing time, and accuracy (based on a labeled subset). Error sources are identified in relationship to trajectory, network, and weather factors. Results indicate poor concordance and widely varying performance, with no algorithm best across all measures. Four high-performing algorithms are identified with at least 90% record-level accuracy; other considerations include accuracy of the inferred number and duration of trips, precision of identified trip end points, and computational resource requirements. Density is a key variable for trip identification in the best-performing algorithms. Proximity to tree canopy, buildings, bridges, and tunnels affects the accuracy of some algorithms more than others. Most algorithms err almost entirely with false trips or false activities, which is a bias of concern for analysis. The importance of trip identification decisions should motivate more thorough reporting to enhance reproducibility and reliability.

1. Introduction

Many cities are promoting the use of active modes of transportation (i.e., walking and cycling) due to expected advantages over other travel modes such as increased physical activity, reduced travel costs, and lack of direct air pollution emissions. At the same time and with similar motivations, a growing body of research seeks better understanding of various aspects of active travel behavior. Conducting detailed empirical analysis of walking and cycling behavior typically involves Global Positioning System (GPS) data. Fortunately, the recent proliferation of fitness and travel smartphone apps, bikeshare systems with location logging, and smartphone-based travel surveys provide an abundance of active travel GPS data.

GPS technologies have been employed in travel surveys since the 1990s, enhancing the detail of travel data collected (Shen & Stopher, 2014). GPS data offer advantages over traditional trip diary methods, such as avoiding under-reporting of trip frequency and over-reporting of trip duration (Houston et al., 2014). GPS data have limitations such as signal loss and signal noise that necessitate processing of the raw GPS data before subsequent analysis.

* Corresponding author.
E-mail addresses: eberjis@mail.ubc.ca (E. Berjisian), alex.bigazzi@ubc.ca (A. Bigazzi).

https://doi.org/10.1016/j.trc.2022.103588
Received 1 June 2021; Received in revised form 16 December 2021; Accepted 27 January 2022
Available online 5 February 2022

Travel GPS data collection methods can be categorized by different levels of engagement by the traveler, which impacts data processing steps. Data collected while a participant goes about their day (passive data collection) creates minimal burden, but typically uses a lower GPS sampling frequency to limit battery power consumption, and generates more non-travel (activity) records (Marra et al., 2019). In contrast, active data collection methods only ask participants to record GPS data during travel, which increases participant burden and may result in incomplete travel data (missed trips), but allows for higher sampling rates and generates primarily travel records - see for example Safi et al. (2014). GPS travel survey participants are often also asked to provide supplementary information about their trips such as mode, trip purpose, or associated activities (Bricka et al., 2012; Shen & Stopher, 2014).

Typical processing steps for raw GPS trajectories are 1) filtering for major and systematic errors, 2) labeling records as being part of trip segments or intervening activities, 3) inferring mode and purpose for trip segments (if needed), and 4) matching GPS records to the street network (map-matching) (Schüssler et al., 2011). Most studies in the literature focus on enhancing techniques for the third and fourth processing steps; considerably less attention has been paid to filtering and trip identification. Both active and passive GPS data collection methods will generate a combination of travel and activity records, and distinguishing the two (i.e., the trip identification step in data processing) can be important for subsequent analysis – particularly microscopic analysis of speed, energy expenditure, etc. This paper aims to determine the best methods for trip identification in the case of walking and cycling GPS data.

2. Literature review

Heuristic trip identification algorithms use a set of rules designed by the analyst to identify trips in GPS trajectories. The rules define thresholds to distinguish trip from activity records using data features such as time difference between records, speed, and record density. For example, many heuristic algorithms use low-speed thresholds in the range of 0 to 0.15 m/s to identify activity records (Dalumpines & Scott, 2017; Rewa, 2012; Schuessler & Axhausen, 2009; Stopher et al., 2008; Tsui & Shalaby, 2006; Wolf et al., 2001). Another common approach for heuristic algorithms is to use a threshold for the dwell time or time difference between consecutive records (sometimes with a speed threshold) to identify activities, typically 120 s in reference to traffic signal cycle lengths (Dalumpines & Scott, 2017; Rewa, 2012; Shen & Stopher, 2013; Stopher et al., 2008; Wolf et al., 2001).

Machine learning techniques have also been used for trip identification – often a version of DBSCAN (Density-Based Spatial Clustering of Application with Noise) (Gong et al., 2018; Hwang et al., 2017; Tran et al., 2013; van Dijk, 2018; Xiang et al., 2016; Yao et al., 2019; Zhou et al., 2017). Similar to heuristic approaches, the machine learning methods employed to date also rely on parameters determined by the analyst. For example, DBSCAN needs a spatial search radius to identify clusters of activity records. Thus, a challenge for both heuristic and machine learning approaches is the selection of appropriate parameter values, which requires either ground-truth data for local calibration or adoption of values from other studies (with uncertain transferability to new datasets or travel contexts).

Independent evaluation of the accuracy of trip identification methods has been limited. Self-reported accuracy (assessed with a variety of methods and reference conditions) ranges from 30% to 100% for heuristic algorithms and 88% to 92% for machine learning methods (Gong et al., 2018; Schuessler & Axhausen, 2009; Stopher et al., 2008; Tran et al., 2013; Wolf et al., 2001; Xiang et al., 2016; Yao et al., 2019). Van Dijk (2018) compared trip identification algorithms using artificial trajectories with different sampling rates and levels of noise. The artificial trajectories were constructed for car, walk, and bike trips by: 1) creating activity records around points of interest (cafés, schools, etc.) and then 2) connecting activity locations by shortest paths (converting polylines to a sequence of discrete records). They found that machine learning methods outperformed heuristic algorithms with accuracy consistently over 91%, compared to heuristic algorithm accuracy between 55% and 80%. All the tested machine learning algorithms required ground-truth training data. Another study compared seven machine learning algorithms trained on ground-truth data to detect activity records as well as transportation modes within GPS trajectories, with activity episode hit ratios of 0.85–1.00 for training and test datasets (Feng & Timmermans, 2016). With the exception of one heuristic algorithm exclusively applied to cycling trips (Rewa, 2012), the existing heuristic and machine learning methods have been applied to multi-modal travel. The lack of specificity to active travel is potentially problematic for algorithm parameters related to speed or other characteristics for which active travel modes systematically differ from motorized modes.

This study is motivated by a key gap in the literature, the lack of independent comparative evaluations of existing trip identification algorithms to assess the transferability of algorithms to the context of active travel. Trip identification methods developed *ad hoc* by researchers can be non-transferable, if heuristics are highly case-specific to the designer's dataset or machine learning methods are over-fit to the training data (Ho et al., 2020). The self-reported accuracy described above was mostly measured with cross-validation methods using a single dataset. Hence, it is essential to test performance on other datasets (external validity) when comparing algorithms. External validation is a crucial step in the generation of scientific knowledge, particularly for machine learning methods and prediction models (Bleeker et al., 2003; Ho et al., 2020). However, external validation studies are rare, contributing to questions about a reproducibility "crisis" in both the natural and social sciences (Baker, 2016; Camerer et al., 2018; Open Science Collaboration, 2015).

Thus, the objective of this study is to conduct a comprehensive and robust evaluation of trip identification algorithms in the literature and make recommendations for the processing of unlabeled active travel GPS data and improving algorithm performance. Algorithms are evaluated on their accuracy, processing time, distinction of trip and activity data characteristics, and interchangeability (the extent to which they produce equivalent trip or activity record labels). Sources of error are examined in relationship to trajectory, network, and weather factors.

3. Method

Fig. 1 illustrates trip and activity records within a GPS trajectory (a timestamped sequence of locations recorded by a GPS device). A basic GPS trajectory contains information on timestamps and geographic coordinates (latitude, longitude) for each record. The study scope includes any algorithm that can be applied to basic GPS data without additional fields (such as the number of tracked satellites or appended sensor data such as from an accelerometer), and published with sufficient information for replication.

To identify existing algorithms, a search of papers published in 2010 through January 2019 was conducted using the Google Scholar search engine with search term combinations of GPS (or global positioning system) with trip end, stop, activity, or point of interest. Additional papers were identified from the reference lists of these papers and peer review. The search yielded 14 trip identification algorithms meeting the study scope (9 heuristic and 5 machine learning). The original algorithms as represented in published material are replicated exactly to the extent possible, including parameter values.

Fig. 2 illustrates the methodology for the comparative evaluation of trip identification algorithms. The algorithms are applied to two real-world travel datasets: a large unlabeled dataset and a smaller dataset with trip or activity labels for each record. A consistent set of basic filtering rules is implemented before applying the trip identification algorithms. Generated trip or activity labels by each algorithm are evaluated based on accuracy of individual record labels (for the labeled dataset), concordance measures to assess agreement (and interchangeability) across algorithms, and discrimination measures representing the distinction between trip and activity data characteristics (as labeled by each algorithm). Algorithm processing times are also compared. Sources of error for each algorithm are identified in relationship to contextual factors through quantitative and qualitative analysis.

3.1. Data

The unlabeled dataset for this study comes from a GPS-based active travel survey conducted in 2017 in metropolitan Vancouver, Canada. Participants were persons of at least 14 years of age who “typically cycle at least once a week”. After obtaining consent, participants completed an online questionnaire that captured demographic information and travel habits and preferences. Participants then recorded one week of their active travel (walking, cycling, running, etc.) using a smartphone application recording GPS-based locations at 1-second intervals. Participants manually started and stopped recording for each trip, also indicating travel mode and trip purpose. Daily email reminders were sent to participants, who were also asked to indicate if any active travel trips were not recorded that day. Recorded GPS trajectories (the sequence of records between the manual start and end of recording) included activity observations and multiple trip segments when participants started recording before the trip and/or failed to end recording immediately at the end of the trip. Only basic GPS data (timestamp and location) were recorded. Most participants used their personal smartphones or other GPS devices to complete the survey, and device details were not recorded.

Recruitment and data collection occurred from June through October 2017. Ten trajectories with transit segments identified by visual inspection and participant comments were removed, as were trajectories recorded outside the metropolitan Vancouver study area. This left 1712 trajectories recorded by 145 individuals, who also reported 129 missed trips (the unlabeled dataset). A separate labeled dataset of 32 trajectories was collected using the same GPS recording methods, but with a prompted recall labeling of records as a trip or activity by one of the study participants. A subset of these trajectories (7) was recorded passively, with the participant instructed to allow the GPS application to run continuously in the background throughout the day, followed by prompted recall of travel mode and purpose for each trip, similar to the actively collected data. Table 1 summarizes the unlabeled and labeled datasets.

Network and weather characteristics were appended to the labeled dataset for error analysis. Street network data for the study area were extracted from Open Street Map (OSM) (OpenStreetMap contributors, 2020). Land cover classification data (tree canopy and buildings) were obtained from the Metro Vancouver open data catalog[1], and assigned to network links based on greatest proportion of intersecting link length. Historical hourly weather data were obtained from an archive of Vancouver International Airport weather station data[2].

3.2. Filtering rules

Uniform filtering rules are applied to all GPS trajectories to control for the effects of filtering on algorithm performance. The set of rules, derived from the literature, is:

1. Delete records with any missing attributes of timestamp or geographic coordinate (latitude or longitude), consistent with Zhou et al. (2017). This rule removes 3% and 9% of records in the unlabeled and labeled datasets, respectively.
2. Identify duplicate timestamps and if present, keep only the earliest one, consistent with Zhou et al. (2017). This rule removes 1% and 0% of records in the unlabeled and labeled datasets, respectively.
3. Reorder or remove reverse time sequences after calculating the time difference between each pair of consecutive records. Removing negative time differences eliminates 10 and 0 records in the unlabeled and labeled datasets, respectively.

[1] http://www.metrovancouver.org/data
[2] https://www.timeanddate.com/weather/canada/vancouver/

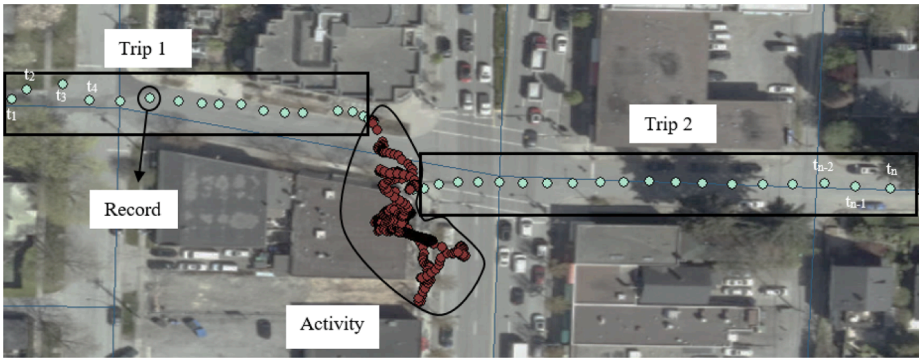

Fig. 1. Illustration of trip and activity records in a GPS trajectory (underlying image from www.arcgis.com).

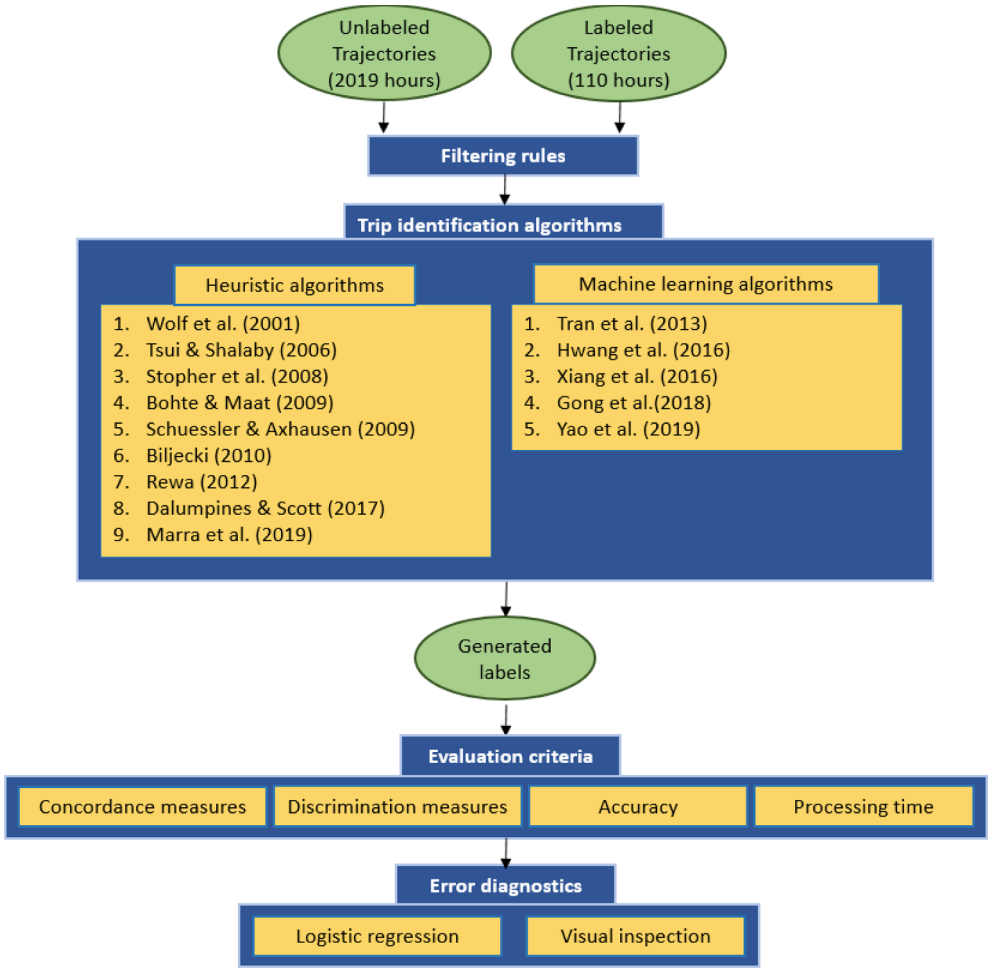

Fig. 2. Illustration of algorithm evaluation method.

4. Remove location errors based on "position jumps" in which the traveled distance is greater than a buffer of 30 m and at a speed of at least 50 m/s (Schuessler & Axhausen, 2009). This rule eliminates 53 and 2 records in the unlabeled and labeled datasets, respectively.
5. Split trajectories at a time gap greater than 8 h, presuming that it represents an activity, consistent with Biljecki (2010). This rule splits 6 and 0 trajectories in the unlabeled and labeled datasets, respectively.
6. Remove trajectories with fewer than 120 records which are unlikely to represent real trips and typically not useful in travel analysis. This rule eliminates 46 and 0 trajectories in the unlabeled and labeled datasets, respectively.

Table 1
Summary of datasets used in analysis.

Feature	Unlabeled data, actively collected	Labeled data, actively collected	Labeled data, passively collected
Trajectories	1712	25	7
Total hours	2019	12	98
Hours per trajectory*	1.2 (18.1)	0.5 (0.3)	14.0 (4.0)
Total records	3,117,348	42,122	280,991
Records per trajectory*	1821 (4646)	1685 (1148)	40,142 (17,062)
Trip proportion of records (versus activity)	Unknown	75%	11%
Travel mode (% of trajectories)			
Bicycle	81%	72%	29%
Walk/run	12%	28%	0%
Both (bicycle and walk/run)	0.2%	0%	71%
Other/unknown	6.8%	0%	0%
Trip purpose (% of trajectories)			
Work/school/errand	72%	88%	71%
Leisure/exercise	12%	8%	0%
Other/unknown/multiple	16%	4%	29%
* mean (standard deviation)			

Cumulatively, this set of filtering rules removes 6% and 9% of the records in the unlabeled and labeled datasets, respectively. The final unlabeled dataset has 2,942,944 records (1653 trajectories, 1219 h) and the final labeled dataset has 295,978 records (32 trajectories, 110 h).

3.3. Heuristic algorithms

Table 2 shows the variables employed in the nine heuristic trip identification algorithms identified in the literature (Biljecki, 2010; Bohte & Maat, 2009; Dalumpines & Scott, 2017; Marra et al., 2019; Rewa, 2012[3]; Schuessler & Axhausen, 2009; Stopher et al., 2008; Tsui & Shalaby, 2006; Wolf et al., 2001). The rules for each algorithm are described in the Supplementary Material, Section S1. The most common variables used to differentiate trip from activity records are dwell time or time gap, distance, and speed (a combination of the two).

3.4. Machine learning algorithms

The five machine learning algorithms identified in the literature all rely on or borrow from DBSCAN, which identifies clusters of records that are located closely together (i.e., have high density). In trip identification, clusters of records are assumed to correspond to activities, and records that do not belong to a cluster are usually classified as trips. Algorithm details are given in the Supplementary Material, section S2. Some algorithms cannot be exactly replicated within the study scope due to extra data requirements, as described below.

3.4.1. Exactly replicated algorithms

The Tran et al. algorithm is a spatial–temporal clustering method based on DBSCAN taking into account noisy records around a stop (Tran et al., 2013). The Hwang et al. algorithm is DBSCAN but instead of random seeds, kernel density estimation is used to refine stop locations to areas with a higher density of records (Hwang et al., 2017). Both the Tran et al. and Hwang et al. algorithms are replicated exactly using the parameter values reported in the original studies.

3.4.2. Partially replicated algorithms

The other three machine learning algorithms could not be exactly replicated within the study scope (using only GPS data). The Xiang et al. algorithm is a sequence-oriented clustering inspired by DBSCAN that captures noise around stationary records and classifies them as stops (Xiang et al., 2016). In the final steps of the algorithm, three criteria are used to correct misidentified labels, one of which relies on geocoded addresses of possible activity locations. This third criterion is excluded from our implementation. The Gong et al. algorithm adds two other constraints (temporal and entropy) to the original DBSCAN to enhance its application for trip identification (Gong et al., 2018). A Support Vector Machine is used in the last step of the algorithm to distinguish non-activity stops (such as at traffic lights), which requires a labeled dataset for model training; this step is excluded from our implementation. The Yao et al. algorithm uses spatial–temporal DBSCAN along with several optimization models to enhance performance (Yao et al., 2019). Two of those optimization models are excluded from our implementation because they require map-matched network link data and traffic congestion data.

[3] Identified through replication in Usyukov (2017).

Table 2
Variables used in heuristic algorithms.

Variable	Wolf et al.	Tsui & Shalaby	Stopher et al.	Bohte & Maat	Schuessler & Axhausen	Biljecki	Rewa	Dalumpines & Scott	Marra et al.
Dwell time or time gap	●	●	●	●	●	●	●	●	●
Distance	●	●	●	●			●		●
Density of records			●		●				●
Heading (azimuth)			●						
Speed		●	●	●	●	●		●	

3.5. Concordance measures

Concordance measures assess the similarity of results across algorithms in terms of 1) the labels for each record, 2) the number of trips within each trajectory, and 3) the distribution of trip durations. Higher concordance indicates higher interchangeability and hence less importance of trip identification algorithm selection to the ultimate analysis results. Table 3 summarizes the concordance measures and associated statistical tests, which were selected to assess inter-rater reliability (treating the algorithms as raters) for the different data types. Cohen's and Fleiss' kappas are used to test pairwise and group agreement, respectively, for record labels as they account for the chance agreement among raters for dichotomous categories (McHugh, 2012; Nichols et al., 2010; Ranganathan et al., 2017). Note that these measures cannot be applied to the Bohte & Maat nor the Biljecki algorithms because those algorithms only split trajectories into trips and do not generate labels for each record. Intraclass correlation coefficient (ICC) is used to test agreement for the generated number of trips, because it applies to numerical "ratings" from multiple raters for a fixed set of observations (trajectories) (Koo & Li, 2016). There is no fixed set of observations for trip duration because each algorithm generated a different number of trips; therefore, agreement among algorithms for trip duration is assessed by comparison of the unmatched distributions of generated trip durations. Similarity of trip duration distributions is tested pairwise between algorithms using Dunn and Kruskal-Wallis statistical tests (Dinno, 2015).

3.6. Discrimination measures

Discrimination measures assess the extent to which each algorithm creates sets of trip and activity records with distinct motion attributes, based on the premise that more discriminating algorithms are more precise. Collectively, trip records are expected to have higher speed, acceleration, and net moved distance, and lower density and heading changes, than activity records. The discrimination measures compare the following distributions for trip and activity records using one-sided Welch t-tests with a 95% confidence level threshold (except for the Bohte & Maat and the Biljecki algorithms which do not generate individual record labels):

1. Instantaneous speed (distance between two consecutive records divided by the corresponding time gap)
2. Acceleration (difference between instantaneous speeds of two consecutive records divided by the corresponding time gap)
3. Record density (number of other records within 50 m and 60 sec)
4. Heading change (between two consecutive records, in degrees)
5. Average event speed (for each set of contiguous trip or activity records)
6. Event net moved distance (between the first and last record in each set of contiguous trip or activity records)

In addition to calculating these six discrimination measures using all labeled records, the first four measures (calculated at the record scale) were also calculated using just the records within 20-sec transition windows at each activity->trip or trip->activity transition reported in the algorithm output, to specifically assess discrimination of precise trip start and end records.

3.7. Accuracy and processing time

For the smaller dataset of labeled records, algorithm-generated labels for each record are compared to identify 1) true trip labels, 2) true activity labels, 3) false activity labels (misidentified as an activity), and 4) false trip labels (misidentified as a trip). From these are

Table 3
Concordance measures.

Feature	Statistical test (95% confidence level threshold)
Dichotomous label of each record (activity or trip)	Fleiss' kappa for all algorithms jointly and Cohen's kappa for pairwise algorithm comparisons
Number of identified trips in each trajectory	Intraclass correlation coefficient (ICC) for 2-way mixed effect with absolute agreement for single measurement for all algorithms jointly and for pairwise algorithm comparisons
Distribution of trip durations for all identified trips	Kruskal-Wallis test and post-hoc Dunn test for pairwise algorithm comparisons

computed four different performance measures: *Accuracy* of all labels $\left(\frac{\text{truelabels}}{\text{alllabels}}\right)$, *Precision* of trip labels $\left(\frac{\text{truetriplabels}}{\text{alltriplabels}}\right)$, *Sensitivity* to trip observations $\left(\frac{\text{truetriplabels}}{\text{truetriplabels}+\text{falseactivitylabels}}\right)$, and *F-score* $\left(2\times\frac{\text{Precision}\times\text{sensitivity}}{\text{Precision}+\text{sensitivity}}\right)$ which is a balance of precision and sensitivity. Processing times across algorithms are compared for a single bicycle trajectory of 69 min (similar to the mean duration in the unlabeled dataset), processed on a desktop computer with 32 Gb of RAM. Not all processing times were recorded and reported because most of the analysis was executed on the Compute Canada multi-core supercomputer. For longer trajectories, the Stopher et al. and Xiang et al. algorithms require more RAM than is available in a typical desktop computer (up to and exceeding 250 Gb).

3.8. Error diagnostics

To investigate sources of error, logistic regression models were estimated for each algorithm using the labeled trajectories (N = 32), with the proportion of falsely labeled records within each trajectory as the dependent variable and trajectory, network, and weather factors as independent variables (retained at a significance threshold of $p < 0.05$). Table 4 lists the independent variables tested in the models and their summary statistics for the labeled dataset. This set of variables was selected after considering a large pool of potential variables, and then refining to eliminate correlation coefficients over 0.7. Network and weather variables were selected based on known sources of GPS errors such as signal obstruction due to tree canopy, multipath error due to tall buildings, and the adverse impact of overcast weather on GPS positioning accuracy (Bricka et al., 2012; Gong et al., 2014; Shen & Stopher, 2014; Yeh et al., 2009). Trajectory variables were included to identify systematic relationships with trajectory characteristics.

The models were estimated using the "stats" package in the statistical software R (R Core Team, 2021). For model interpretation, elasticity is used to represent the impact of a 1% change in each variable on the error rate percentage. Aggregate point elasticity is computed for all continuous variables, and arc elasticity for the binary variable "Cloudy" (Hensher et al., 2005). Falsely labeled records were also mapped for each trajectory and visually inspected for supporting evidence.

4. Results

Due to the limitation on RAM, Stopher et al. failed to run for 2 trajectories in the unlabeled dataset (with durations of 14.2 and 29.3 h). All algorithms ran successfully for the 25 trajectories in the actively collected labeled dataset, but 1 trajectory (with duration of 20.6 h) from the passively collected labeled dataset failed to complete for Stopher et al. due to excessive time needed on Compute Canada servers (more than 28 days, the available allocation). Xiang et al. failed to complete on both the unlabeled and passively collected labeled datasets due to similarly excessive time needed (more than 28 days each).

4.1. Concordance measures

Fig. 3 gives the shares of trip and activity record labels generated by each algorithm for the unlabeled and labeled datasets. The uneven bar heights are due to some algorithms failing to run on certain long trajectories, as described above. In addition, the Bohte & Maat algorithm eliminates records it identifies as "garbage points", which removed 90% and 99% of records in the unlabeled and labeled datasets, respectively.

The concordance measures indicate substantial inconsistency in labels among algorithms. Just 1% of records have the same label assigned by all (successfully executing) algorithms for the unlabeled dataset. Fleiss's kappa of 0.04 for the unlabeled dataset and 0.07 for the labeled dataset indicate slight agreement among algorithms collectively. Pairwise comparisons by Cohen's kappa (given in Supplementary Material) confirm the generally poor agreement in labels between algorithms, with more similar labels within the

Table 4
Trajectory, network, and weather variables for each trajectory in the labeled dataset, used in error diagnostic models.

Variable	Definition	Average (standard deviation)
Walk proportion	Share of walk/run records to total records (%)	22 (42)
Speed	Arithmatic mean of instantaneous speed for each record, as defined above (m/sec)	3.61 (2.19)
Density	Arithmatic mean of the number of records within 50 m and 60 sec of each record	1856 (4566)
Activity center spacing	Arithmatic mean of distance between center of consecutive activity segments (m)	5073 (4355)
Missing data proportion	Duarion of missing seconds (time gap > 1 sec) divided by total duration	0.11 (0.14)
Bridges and tunnels	Total length of links with OSM keys bridge and tunnel intersecting (completely or partially) a 50-meter buffer around the trajectory divided by the total length of links intersecting the buffer	0.04 (0.06)
Tree cover	Total length of links with tree canopy land cover intersecting (completely or partially) a 50-meter buffer around the trajectory divided by the total length of links intersecting the buffer	0.32 (0.10)
Buildings	Total length of links with building land cover intersecting (completely or partially) a 50-meter buffer around the trajectory divided by the total length of links intersecting the buffer	0.08 (0.04)
Cloudy	Binary variable for cloudy weather for majority duration of trajectory recording, based on descriptive status of: "cloudy", "fog", "(mostly) rainy", or "overcast"	Cloudy = 8 (25%)

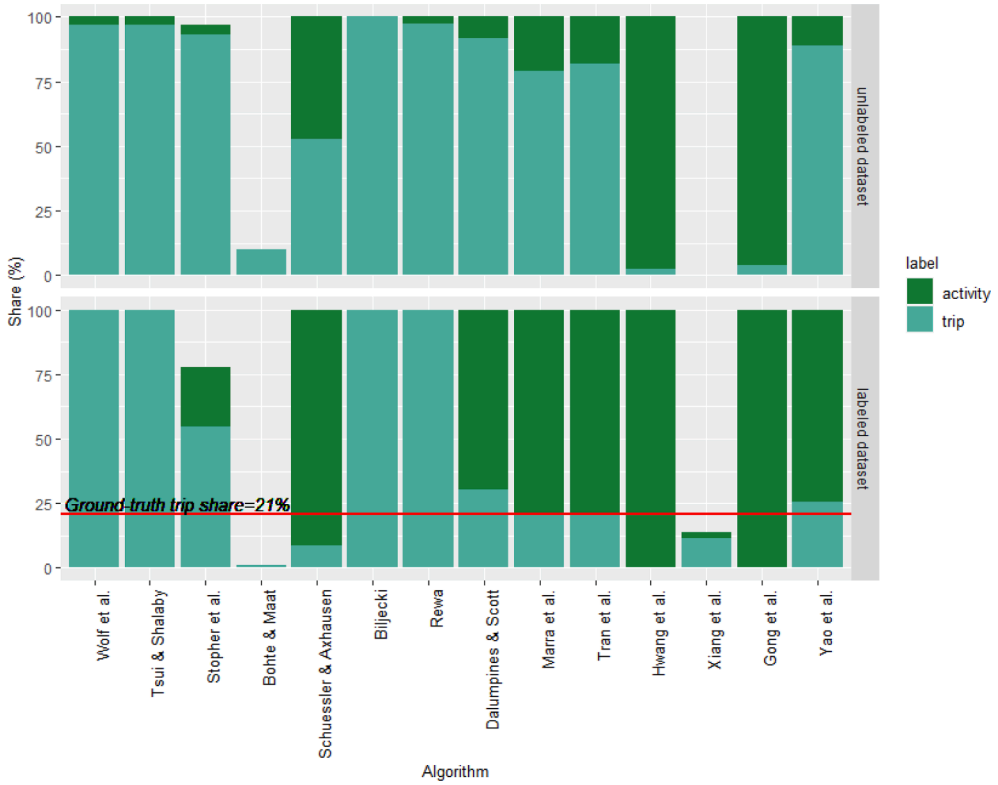

Fig. 3. Shares of trip and activity labels generated by each algorithm for the unlabeled and labeled datasets.

heuristic and machine learning categories than across them (excepting the substantial agreement between Tran et al. and Marra et al.). The unlabeled dataset was collected actively, and so the majority of records are expected to be trip observations – which is inconsistent with results from several of the algorithms, especially Gong et al. and Hwang et al. For the labeled dataset, the ground-truth share of trip labels is also inconsistent with the shares reported by most of the algorithms.

Fig. 4 gives the total number of trips generated by each algorithm for each dataset, which varies substantially for some algorithms. The number of trips per trajectory also varies substantially across algorithms (averaging from less than 1 to 7). The ICC values for number of trips per trajectory are 0.05 and 0.00 for the unlabeled and labeled datasets, respectively, which indicates poor collective agreement. Pairwise comparisons also indicate poor agreement between algorithms in the number of trips identified (see Supplementary Material section S3). Schuessler & Axhausen and Gong et al. for the unlabeled dataset and Dalumpines & Scott for the labeled dataset greatly over-partition trips compared to the other algorithms.

Fig. 5 displays the distribution of trip durations for trips identified by each algorithm. In the labeled dataset, median trip durations for the heuristic algorithms range from 1.9 (Dalumpines & Scott) to 30.1 (Biljecki) minutes (average of 19.0 and standard deviation of 11.7 across algorithms) while for the machine learning algorithms they range from < 0.1 (Gong et al.) to 21.6 (Xiang et al.) minutes (average of 6.4 and standard deviation of 10.3 across algorithms). The ground-truth median trip duration for the labeled dataset was 13.0 min. Kruskal-Wallis tests indicate a significant difference among trip duration distributions for both datasets (chi-square of 19,325 and 290 for unlabeled and labeled datasets, with 13 degrees of freedom). Dunn tests on pairwise comparisons of algorithms also indicate significant differences among trip duration distributions for most of the algorithms – particularly for the unlabeled dataset (see S3 within Supplementary Material).

Collectively, the three concordance measures reveal substantial inconsistency in individual record labels and identified trip characteristics (frequency and duration) across algorithms. Among machine-learning algorithms, Gong et al. and Hwang et al. tend to label almost all records as activities. The heuristic algorithms tend to identify more records as trips. The Schuessler & Axhausen, Dalumpines & Scott, and Gong et al. algorithms deviate most from the other algorithms in terms of identifying more frequent and shorter trips in the data (although not consistently for both datasets).

4.2. Discrimination measures

Ideally, an algorithm will not only identify the occurrence of trips and activities, but also precisely mark the transition points between them so that no activity observations are included in the subsequent travel analysis. Fig. 6 illustrates a sequence of generated labels for the end segment of an example bicycle trajectory in the labeled dataset (only algorithms identifying both labels for this trajectory are included in the figure). None of the algorithms identify the exact start of the trip; Dalumpines & Scott and Marra et al.

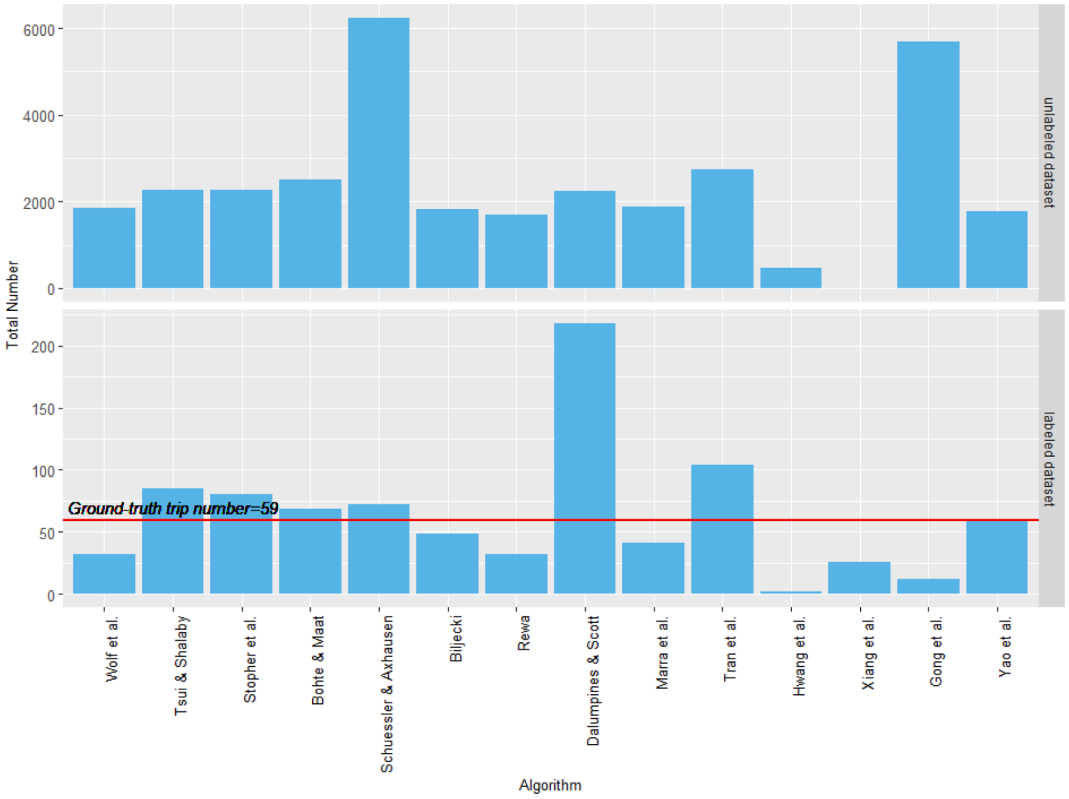

Fig. 4. Total number of trips generated by the algorithms for unlabeled and labeled datasets.

include activity records in the trip, while Tran et al., Xiang et al., and Yao et al. include trip records in the activity cluster. Gong et al. performs worst as it includes all trip records shown in Fig. 6 in the activity. Xiang et al. comes closest to identifying the correct trip endpoint, missing less than a block of trip observations.

Table 5 summarizes the significant discrimination measures by algorithm, including the 6 overall measures (instantaneous and average speed, acceleration, record density, heading change, and net moved distance) and the 4 transition-window measures (instantaneous speed, acceleration, record density, and heading change). Complete results are reported in S4 of the Supplementary Material. Six of the algorithms produced homogenous labels (all trip or all activity) for at least one of the datasets, in which cases the discrimination measures could not be calculated. Schuessler & Axhausen and Marra et al. generated labels with all significant discrimination measures for both datasets (all the discrimination measures for the ground-truth data were also significant, supporting the relevance of the discrimination measures). Most other algorithms generated 7 to 9 significant discrimination measures out of the 10 (if they could be calculated), with the exception of Hwang et al., which failed to generate distinct trip and activity datasets by most of the measures. The transition-window discrimination measures were less often significant, particularly record density and heading change which were significant for 4 and 3 of the 14 algorithms in the unlabeled and labeled datasets respectively.

4.3. Accuracy

Accuracy of aggregate data features is included in the previous results comparing algorithm-generated trip/activity shares (Fig. 3), trip frequency (Fig. 4), and trip duration (Fig. 5) with ground-truth data. Fig. 7 gives accuracy of individual record labels by each algorithm for the labeled dataset, divided between actively and passively collected data. The algorithms overall had higher accuracy for the actively collected data. Several heuristic algorithms failed to identify any activity observations in either dataset. Of the machine learning algorithms, Hwang et al. failed to identify any trip observations while Gong et al. only identified 55 trip records (out of 62,401 true trip records). Dalumpines & Scott and Stopher et al. tended to over-label as trips, while Schuessler & Axhausen tended to over-label as activities. Marra et al. and Dalumpines & Scott had the highest accuracy among the heuristic algorithms: Marra et al. generated 89% and 96%, correct labels for the actively and passively collected data, respectively, and Dalumpines & Scott generated 91% and 91%, respectively.

Machine learning algorithms were generally more accurate than the heuristic algorithms, with 77% average accuracy versus 57% for the heuristic algorithms. Tran et al. had the highest accuracy among the machine learning algorithms (91% and 99% correct labels for the actively and passively collected data, respectively), followed by Yao et al. (90% and 95% correct labels for the actively and passively collected data, respectively). Xiang et al. had similar performance on the active dataset but failed to complete execution on

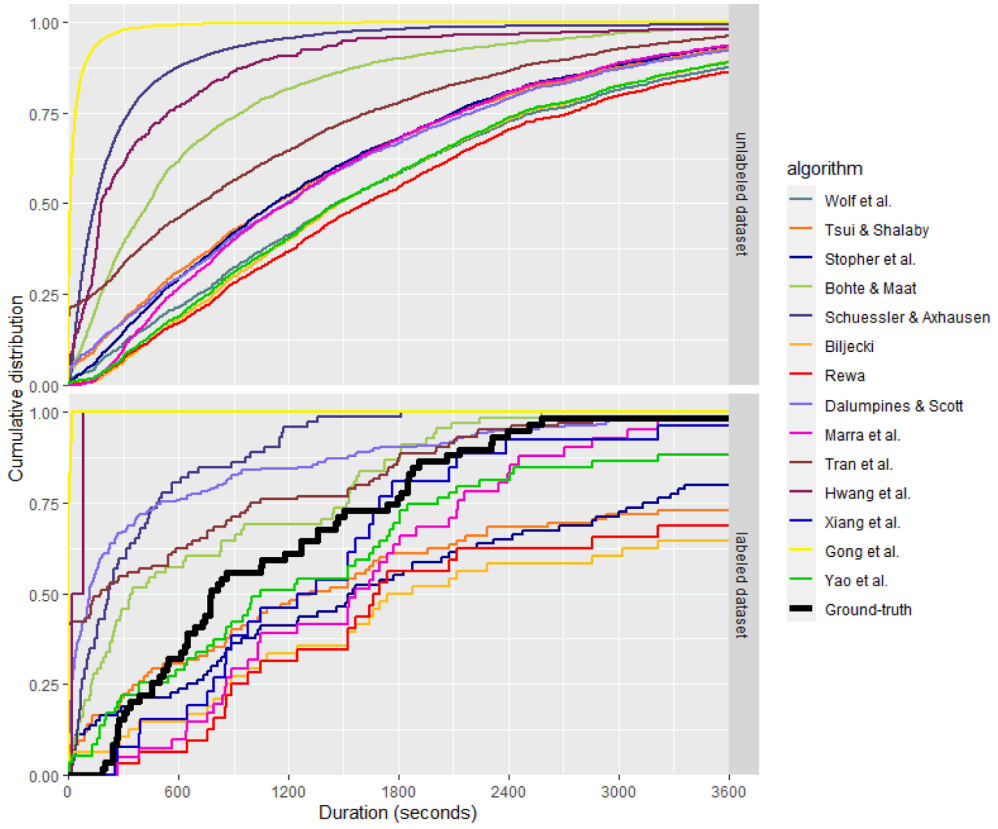

Fig. 5. Cumulative distributions of identified trip durations (up to one hour).

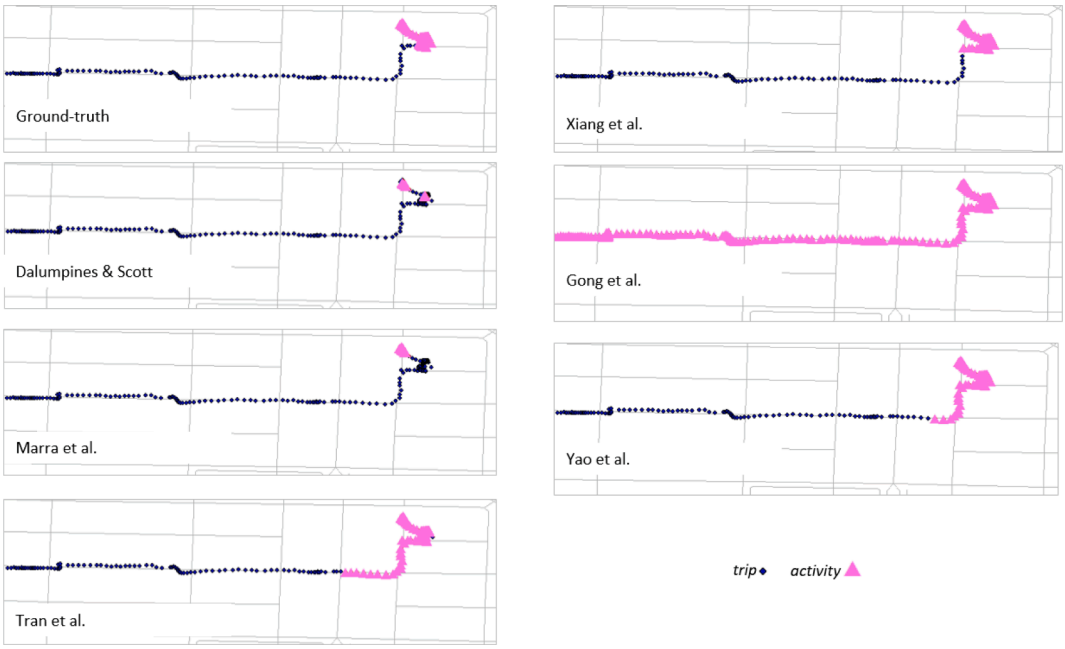

Fig. 6. Sequence of generated labels for the end segment of an example bicycle trajectory.

Table 5
Number of significant discrimination measures (out of 10) for each algorithm.

Algorithm	Number of significant (at 95% confidence) discrimination measures (out of 10)	
	Unlabeled dataset	Labeled dataset
Wolf et al.	7	NA*
Tsui & Shalaby	7	NA*
Stopher et al.	8	8
Bohte & Maat	NA*	NA*
Schuessler & Axhausen	10	10
Biljecki	NA*	NA*
Rewa	7	NA*
Dalumpines & Scott	9	9
Marra et al.	10	10
Tran et al.	10	7
Hwang et al.	3	2
Xiang et al.	NA**	9
Gong et al.	9	8
Yao et al.	8	6

* no comparison possible because all records labeled the same (as all trips or all activities)
** failed to complete execution due to extensive processing time

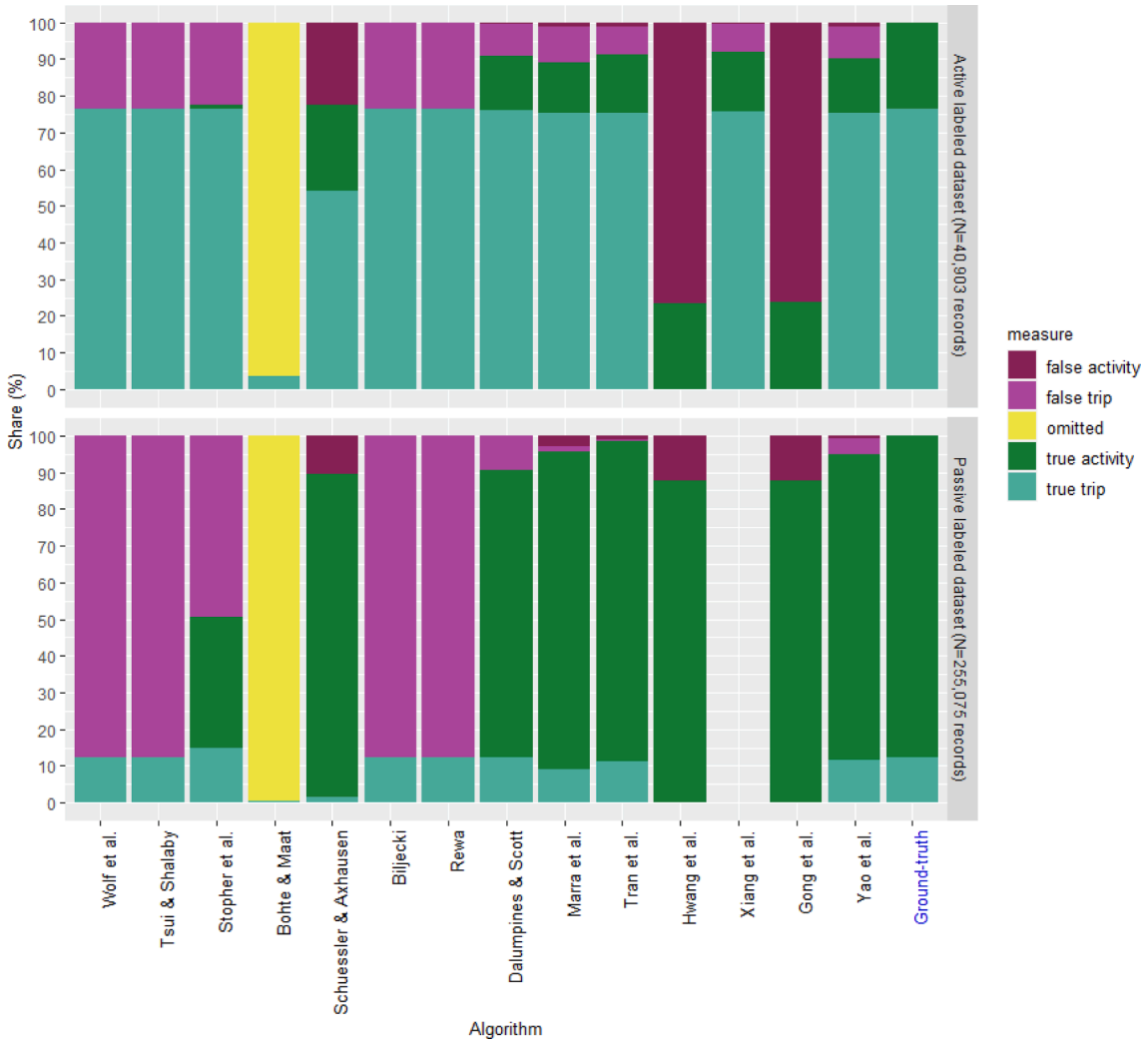

Fig. 7. Algorithm accuracy for the actively and passively collected datasets.

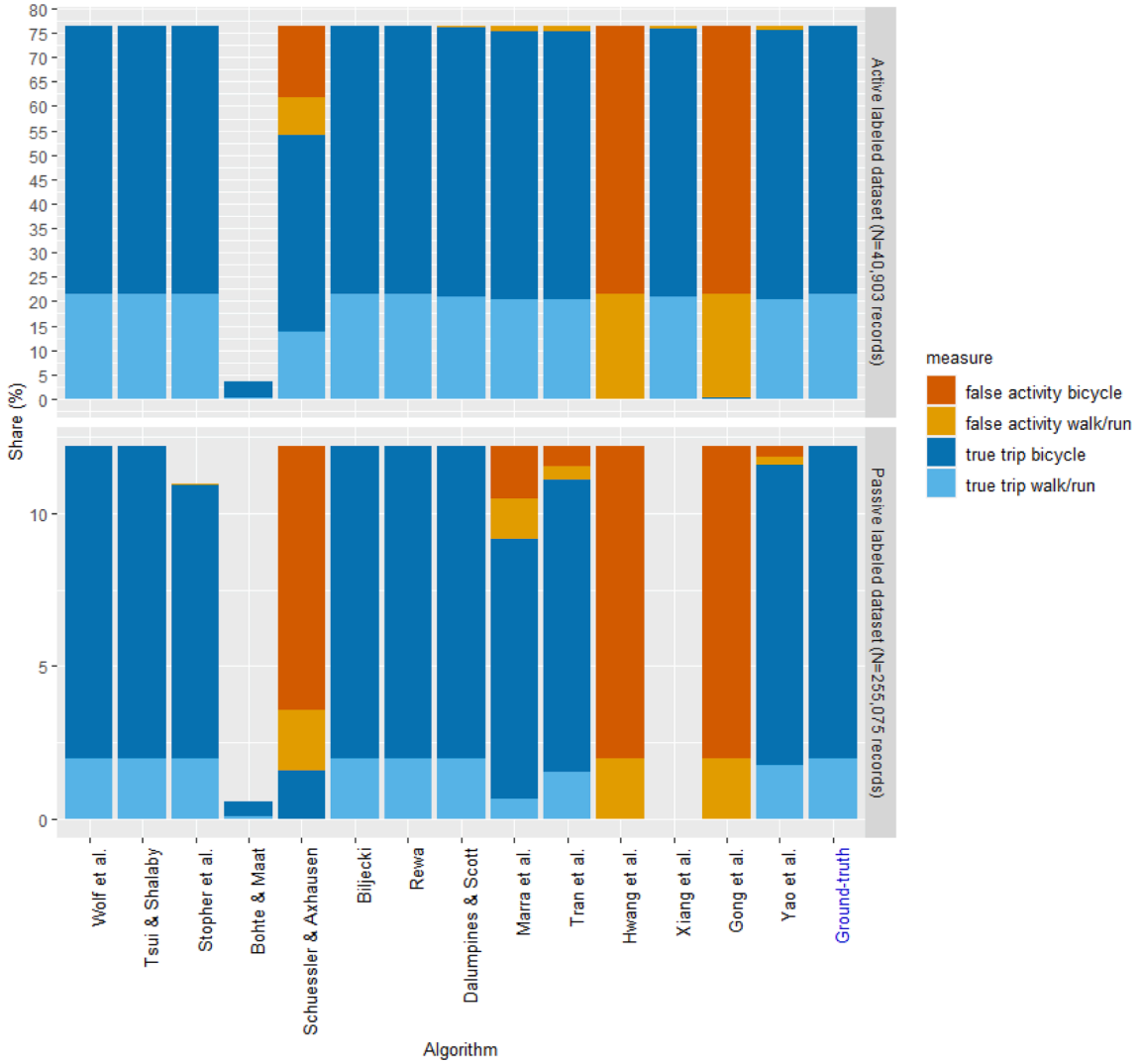

Fig. 8. Algorithm accuracy with respect to travel mode for trip records.

the passive dataset after 28 days on the supercomputer.

Fig. 8 gives algorithm accuracy by travel mode for observations of walking and cycling in the labeled dataset (constituting 76% of the actively collected and 12% of the passively collected data). Schussler & Axhausen and Marra et al. are the only heuristic algorithms with substantial false activity labels, produced at a higher error rate for walk/run trips than bicycle trips. The highest accuracy machine learning algorithms, Tran et al. and Yao et al., both also have higher error rates for walk/run observations.

Fig. 9 gives accuracy and F-score measures for each algorithm, with accuracy representing the overall correctness of all labels and F-score representing the ability of the algorithm to produce reliable trip labels. Due to the tendency of algorithms to err on the side of either trip or activity labels, the two performance measures provide different perspectives on algorithm performance. This difference is particularly dramatic for several algorithms with high accuracy in the passively collected dataset (Schuessler & Axhausen, Hwang et al., Gong et al., Yao et al.). Tran et al., followed by Marra et al. and Dalumpines & Scott, have the most consistent performance between measures. Higher F-scores indicate better reliability of the algorithms in processing data for analysis of travel than of activities (particularly for actively collected data), and vice versa. All four accuracy performance measures are given for each algorithm in the Supplementary Material.

Due to the high performance of the Tran et al. algorithm, four modifications are considered with the aim of further increasing the algorithm's accuracy for walking and cycling trips. The original algorithm uses an Eps (spatial search radius) of 100 m with a MinTime (minimum temporal duration) of 5 min to cluster records into a stop with a K (allowed noisy records) of 3 noisy records outside of the stop cluster. These parameter values are recalibrated to 50 m, 3 min, and 4 records, respectively, based on trials of the labeled dataset. A second modification is labeling the first and last records, and any record whose neighborhood (EKN – see S2.3 in the Supplementary Material) contains the first or the last record, as a core point regardless of the other conditions. This modification increases the

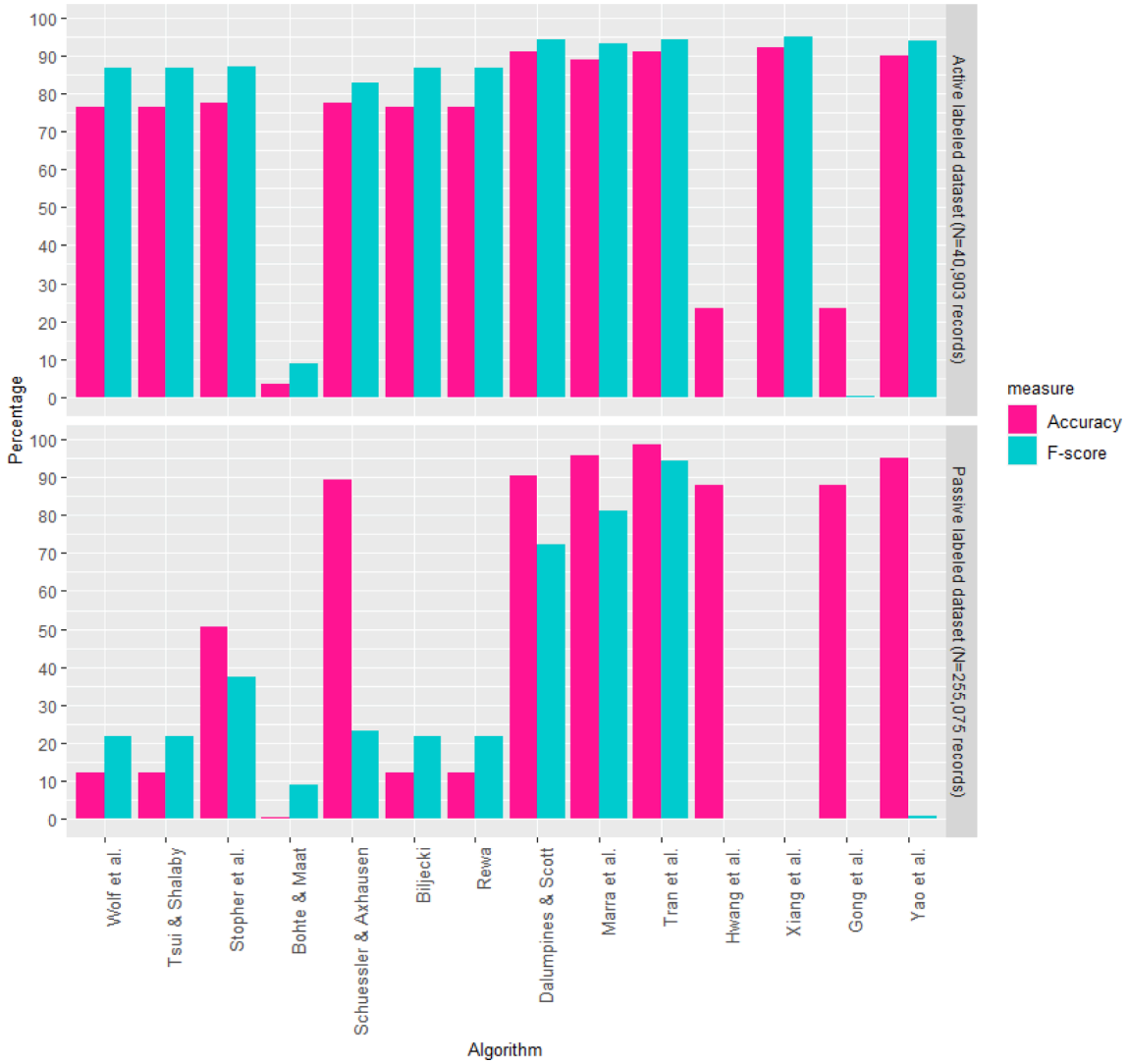

Fig. 9. Accuracy and F-score measures for each algorithm.

likelihood of the first and last records being labeled as activities, based on inspection of the labeling results. The third modification controls empty neighborhoods (EKN) to have an average speed of less than 0.56 m/s (lowest walking speed threshold based on Biljecki (2010)) and consequently label such neighborhoods as a core point. This modification reduces the likelihood of walking episodes in signal loss to be falsely labeled as activities. The final modification merges close activities to avoid false trips within an activity. Together, the four modifications lead to improved accuracy of 97% and 99% for the active and passive datasets, improvements of 6% and 0.4% from the original parameters. Similar modifications/re-calibrations were explored for the other machine learning algorithms as well; accuracy somewhat improved (within 10%), but the comparative performance among algorithms was not substantially altered (and was not validated with a separate set of data).

4.4. Processing time

The processing times for the sample trajectory (with 4100 records) on a desktop computer ranged from 1 s to 113 h across the algorithms. Xiang et al. was an outlier at 4.7 days; the average of the other algorithms was just 0.4 min. Xiang et al. was also the only algorithm with a processing time longer than the data duration (93 sec processing per sec of data). All other algorithms had processing times of 0.05 sec per sec of data or less. Other than Xiang et al., machine learning algorithms had a comparable processing time (average of 27 sec, ranging 16–50) to the heuristic algorithms (average of 36 sec, ranging 1–207). Algorithms using density variables had longer processing times; particularly Marra et al. and Schuessler & Axhausen at 207 and 87 sec, respectively.

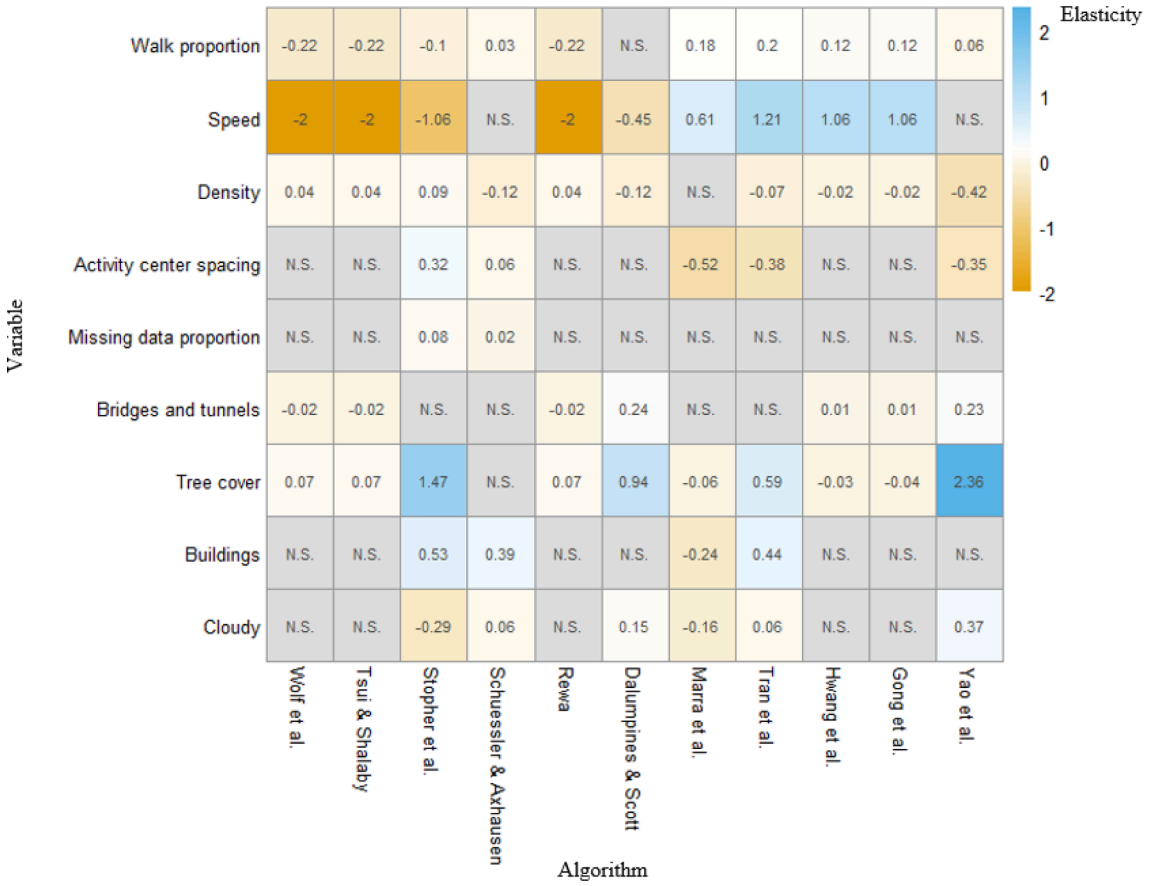

Fig. 10. Elasiticity values for the effects of explanatory variables on algorithm error rates; N.S. indicates not significant at p less than 0.05.

4.5. *Error diagnostics*

Fig. 10 gives error diagnostic (logistic regression) model elasticity results for each algorithm (except Bohte & Maat and Biljecki which generated only one label type, and Xiang et al. which failed to complete on a large share of the data). Positive elasticity indicates an increase in algorithm error with an increase in that variable. Error rates for some algorithms increased substantially with trajectory proximity to tree cover, and to a lesser extent with proximity to buildings, bridges, and tunnels, likely due to signal blocking, multipath, and urban canyon effects on GPS location accuracy (Bricka et al., 2012; Gong et al., 2014; Shen & Stopher, 2014). Cloudy weather significantly increased error rates for four of the algorithms, also likely due to degraded GPS location accuracy (Yeh et al., 2009). Missing data proportion only significantly increased error rates for Stopher et al. and Schuessler & Axhausen – and in both cases with a relatively small elasticity – indicating that most algorithms handle missing data relatively well (likely better than GPS location errors). Some elasticities in Fig. 10 have the opposite of the expected sign (a beneficial effect of tree cover, indicated by negative elasticity, for example), but these also tend to have small magnitude, and are likely due to correlation with another factor (the inverse relationship between tree cover and street canyons, for example).

Some differences in parameter signs between algorithms relate to each algorithm's dominant error direction. The accuracy of algorithms that err toward false activities (Schussler & Axhausen, Gong et al., and Hwang et al.) tend to be similarly affected by trajectory characteristics (i.e., walk proportion, speed, density), while the accuracy of algorithms that err toward false trips (Wolf et al., Tsui & Shalaby, Stopher et al., Dalumpines & Scott, Rewa, and Yao et al.) tend to be oppositely affected by the same trajectory characteristics. Marra et al. and Tran et al. have more balanced errors between trips and activities, but their accuracy is still affected by trajectory characteristics. Error rates generally decreased with speed for algorithms that use speed or distance in their heuristic, while error rates generally decreased with density for algorithms using DBSCAN. Most density-based algorithms had increased error rates for trajectories with a higher proportion of walking records, whereas most other algorithms had the opposite. Three of the highest-accuracy algorithms (Marra et al., Tran et al., and Yao et al.) performed better with activity centers spaced further apart, whereas two others (Stopher et al. and Schuessler & Axhausen) performed worse.

Further error diagnosis at the record level was undertaken by visual inspection of maps of record label accuracy by each algorithm. All the heuristic algorithms except Schussler & Axhausen fail to identify short (<3 min) activity segments at either end of actively collected trajectories, due to minimum activity duration thresholds. In addition, zero speed thresholds for activities in Wolf et al., Tsui

& Shalaby, Stopher et al., and Rewa prevent accurate identification of activity records with location noise. Dalumpine & Scott avoids this problem with a non-zero activity speed threshold, but still fails to identify some activities with high location noise. Schussler & Axhausen uses a different approach with a low-threshold density heuristic that tends to produce false activity records – particularly for slow (walking) trips. The density-based rule in Stopher et al. fails to identify activities due to a small spatial search radius. Conversely, Marra et al. allows for longer and larger activity clusters which leads to higher accuracy. False activities generated in the passive dataset are mostly in the transition window between trips and activities.

Similar to the heuristic algorithms, Tran et al., Xiang et al., and Yao et al. fail to identify short activity segments at either end of actively collected trajectories. Tran et al. also tends to include slow trip records in nearby activity clusters, and noisy activity records in adjacent trips or as false trips (if the number exceeds the threshold of allowed noisy records). Hwang et al. and Gong et al. also tend to create false trip labels due to a tight spatial search radius and low minimum number of records required to form an activity cluster. Also, the minimum number of records required for an activity in both algorithms leads to missed activities when signal loss occurs during an activity. Yao et al. reduces these problems with a larger spatial search radius and ensuring adequately long minimum duration for activities.

5. Discussion and conclusions

The results overall show that the 14 evaluated trip identification algorithms have dramatically varying performance for active travel data. The concordance measures indicate poor agreement among the algorithms in terms of individual record labels (kappa statistics of less than 0.1, and just 1% of records labeled the same by all algorithms), the number of inferred trips (averaging 1 to 7 trips per trajectory by different algorithms), and the duration of the inferred trips (averaging 1 to 30 min per trip by different algorithms). On the practical side, processing times range from seconds to days per GPS trajectory for different algorithms, although most executed in under 0.05 s processing time on a desktop computer per second of data. Two of the algorithms encountered RAM limitations for long trajectories on a desktop computer (even on a 250 Gb supercomputer), which likely makes them infeasible for use in many analyses – particularly on passively collected GPS data.

Due to the widely varying performance, selection of a trip identification algorithm is an important decision for walking and cycling GPS data processing, with substantial implications for subsequent analysis. Tran et al. had the highest accuracy for the labeled data: 91% and 99% for the actively and passively collected data, respectively. Three other algorithms achieved accuracy of around 90% for both actively and passively collected data: Yao et al., Dalumpines & Scott, and Marra et al. Machine learning algorithms were generally more accurate than heuristic algorithms (particularly for passively collected data), supporting the results reported by van Dijk (2018) but with real-world data. Accuracy was higher for cycling trips than trips by foot (walk or run), likely because cycling trips are at higher speeds and so easier to distinguish from activities. With a few exceptions, accuracy was also higher for the actively collected than passively collected data, due to a tendency to over-label as trip observations.

Overall record-level accuracy is not the only consideration for selecting an algorithm, however. False activity versus false trip labels may have different effects on subsequent analysis, and the preferred algorithm would depend on the study objectives (e.g., microscopic versus macroscopic travel features). For example, Tran et al. and Marra et al. had higher record-level accuracy, but Yao et al. more accurately inferred the number of trips in the labeled dataset. Dalumpines & Scott also yielded erroneously high trip frequency and short duration in the labeled dataset. Errors for the highest-accuracy algorithms were almost entirely false trips (to a lesser extent for Tran et al. and Marra et al.), meaning they do not miss any trip observations, but do include activity observations in the inferred trips. False trip labels would tend to bias subsequent estimates of travel speed and energy expenditure (low) – and vice versa for false activity labels if the lowest-speed trip records are mislabeled. All algorithms erred almost entirely on the side of trips or activities, particularly for the actively collected dataset. The implications of potentially non-random, asymmetrical errors in trip identification are important to consider when interpreting and reporting results from GPS-based analyses.

The discrimination measures provide another perspective on algorithm performance, particularly the transition window measures which indicate the precision of the identified trip start and end points. Only the discrimination measures for Schuessler & Axhausen and Marra et al. were universally significant. Tran et al. and Yao et al. had high accuracy, but less than half of their transition window discrimination measures were significant, suggesting they may be less precise in distinguishing trip ends. Dalumpines & Scott fared better, with mostly significant transition window discrimination measures (all except heading change).

Other considerations beyond accuracy, error type, and trip end precision include computational power (only limiting for a few algorithms), complexity of implementation (generally easier for heuristics), and requirements for parameter selection or calibration data. All factors combined, the Tran et al. algorithm is likely to be the best to use for trip identification in active travel GPS trajectories, with the caveat that it tends to produce false trips within activities and may over-estimate the number of trips. It also requires calibration of three parameters in advance (spatial search radius, minimum temporal duration for activity clusters, and maximum noisy points allowed around activity records). Marra et al., Yao et al., and Dalumpines & Scott are competitive alternatives which may be preferred in some contexts. For reference, Tran et al. required 10 h to process almost 3 million records in the unlabeled dataset. The algorithm can be refined for active travel data as described above, with modest improvements in accuracy – although the transferability of the modifications to other datasets is unknown.

The original papers describing trip identification algorithms reported accuracy of 30% to 100%, although most did not report both false trip and false activity rates (which is important because most err completely on one side or the other). Our external validation finds that the heuristic algorithms except Stopher et al. underperform the initially reported accuracy (when reported), likely because of poor transferability of parameters/thresholds to different travel modes, GPS devices, and/or data resolution. The machine learning algorithms all originally reported accuracy of 88%-92%, consistent with our validation results for the actively collected dataset (other

than Hwang et al. and Gong et al.). The performance of Xiang et al, Gong et al., and Yao et al. may have suffered from the partial implementations described above.

The accuracy of all algorithms varies with trajectory, network, weather, and possibly other characteristics, which should be considered in application. For example, performance tended to be worse for trips in proximity of buildings, bridges, tunnels, and tree cover, and for passively collected data. Error diagnostics also indicate likely approaches to improve trip identification algorithm performance on walking and cycling GPS data. Most algorithms fail to mark short duration activities at the trajectory ends, which could be corrected by relaxing activity duration thresholds at trajectory ends in actively collected data. Speed thresholds for activities should be higher than zero – but this comes at the cost of falsely labeling slow-moving trip observations (as is the case in Dalumpines and Scott).

Density appears to be the key feature for accurate trip identification in walking and cycling GPS data, and should be considered in future heuristic algorithms. The most successful heuristic and machine learning algorithms (Marra et al. and Tran et al.) are based on flagging activities by their density with sufficiently wide spatial and temporal radii (250 m with 600 min, and 100 m with 300 min). These algorithms also use lower thresholds than others for the minimum number of records in an activity cluster (allowing activities of just 1 or 2 records in the presence of signal loss). Another aspect that appears to enhance algorithm performance is allowing noisy records within an activity, accomplished in Marra et al. by merging short trips between activities and in Tran et al. by allowing up to three noisy records to fall outside of an activity cluster. The drawback is that these algorithms also tend to include slow-moving trip records within a nearby activity (particularly problematic for walking trips). This type of systematic error could create bias in the processed dataset by systematically reducing observations of trips by slower travelers (elderly pedestrians or less-confident cyclists, for example).

A strength of this study is the external validation of existing algorithms using independent GPS data from real travel. Algorithm performance on other data sets, from other regions and collected using different (non-smartphone) GPS devices, may vary, and should be investigated in future research with different validation data. Beyond the factors investigated above (mode, speed, missing data, land cover, weather, etc.), algorithm performance may vary with other factors such as GPS device accuracy, precision, or resolution. Algorithm performance may be improved with a customized set of filtering rules (such as gap-treatment), rather than the consistent set used for all algorithms in this study in order to isolate algorithm performance. Algorithm performance may also be enhanced with ancillary GPS data fields or datasets (accelerometers, etc.), which we excluded to replicate the condition of common datasets.

A significant challenge for all methods is the amount of ground-truth (labeled) training data needed for development, validation, and calibration. A range of practical constraints related to data access, privacy policies, ethical considerations, and computational resources impedes the ability of researchers to utilize multiple datasets for studies on GPS data processing methods. For examples, 12 of the 14 papers presenting trip identification methods evaluated in this paper used data from a single city, as did both of the papers that previously compared machine learning trip identification methods (one of which used simulated data) (Feng & Timmermans, 2016; van Dijk, 2018). Collective action toward a privacy-protecting model of disaggregate data sharing would enable more robust validity tests in all studies, and should be pursued. In addition, because testing transferability is cost-prohibitive for individual analyses focused on data applications rather than data processing methods, future work on algorithm development should focus on generalizability of the methods, with minimal reliance on context-specific variables or local training data and external validity testing as part of the primary assessment of algorithm performance.

In summary, trip identification is a crucial step in GPS data processing that has received insufficient attention. Processing large GPS datasets requires selection of a trip identification method and associated parameter values, and yet many papers using GPS data neglect these details in their methodology description. Our findings show that selection of a method will have a substantial impact on the processed data, in terms of which observations end up in the “trip” and “activity” sets, and the inferred macroscopic travel characteristics (number and frequency of trips). Two machine learning algorithms (Tran et al. and Yao et al.) and two heuristic algorithms (Marra et al. and Dalumpines & Scott) outperformed the others for our validation data, with some caveats. Future research should examine the impacts of other processing decisions (e.g., filtering) and data features (e.g., temporal resolution) on processing accuracy and subsequent analysis. We hope that illustrating the importance of trip identification decisions will spur more thorough reporting of full data processing details to enhance reproducibility and reliability.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank WestGrid (www.westgrid.ca) and Compute Canada (www.computecanada.ca) for providing computational support. We also thank the study participants as well as Amr Mohamed and other members of the REACT Lab at the University of British Columbia for assistance in data collection and feedback. This research was enabled by support from Social Science and Humanities Research Council of Canada (SSHRC) Insight Development Grant #430-2019-00049.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.trc.2022.103588.

References

Baker, M., 2016. 1,500 scientists lift the lid on reproducibility. Nature News 533 (7604), 452–454.

Biljecki, F., 2010. Automatic segmentation and classification of movement trajectories for transportation modes. Delft University of Technology.

Bleeker, S.E., Moll, H.A., Steyerberg, E.W., Donders, A.R.T., Derksen-Lubsen, G., Grobbee, D.E., Moons, K.G.M., 2003. External validation is necessary in prediction research: a clinical example. J. Clin. Epidemiol. 56 (9), 826–832.

Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. Transport. Res. C: Emerg. Technol. 17 (3), 285–297.

Bricka, S.G., Sen, S., Paleti, R., Bhat, C.R., 2012. An analysis of the factors influencing differences in survey-reported and GPS-recorded trips. Transport. Res. C: Emerg. Technol. 21 (1), 67–88.

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B.A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., Isaksson, S., Manfredi, D., Rose, J., Wagenmakers, E.-J., Wu, H., 2018. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. Nat. Hum. Behav. 2 (9), 637–644.

Dalumpines, R., Scott, D.M., 2017. Making mode detection transferable: extracting activity and travel episodes from GPS data using the multinomial logit model and Python. Transport. Plan. Technol. 40 (5), 523–539.

Dinno, A., 2015. Nonparametric pairwise multiple comparisons in independent groups using Dunn's test. Stata J. 15 (1), 292–300.

Feng, T., Timmermans, H.J.P., 2016. Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data. Transport. Plan. Technol. 39 (2), 180–194.

Gong, L., Morikawa, T., Yamamoto, T., Sato, H., 2014. Deriving personal trip data from GPS data: A literature review on the existing methodologies. Procedia-Soc. Behav. Sci. 138, 557–565.

Gong, L., Yamamoto, T., Morikawa, T., 2018. Identification of activity stop locations in GPS trajectories by DBSCAN-TE method combined with support vector machines. Transp. Res. Procedia 32, 146–154.

Hensher, D.A., Rose, J.M., Rose, J.M., Greene, W.H., 2005. Applied choice analysis: a primer. Cambridge University Press.

Ho, S.Y., Phua, K., Wong, L., Goh, W.W. Bin, 2020. Extensions of the External Validation for Checking Learned Model Interpretability and Generalizability. Patterns, 1 (8), 100129.

Houston, D., Luong, T.T., Boarnet, M.G., 2014. Tracking daily travel; Assessing discrepancies between GPS-derived and self-reported travel patterns. Transport. Res. C: Emerg. Technol. 48, 97–108.

Hwang, S., Evans, C., Hanke, T., 2017. Detecting stop episodes from GPS trajectories with gaps. In: Seeing Cities Through Big Data. Springer, pp. 427–439.

Koo, T.K., Li, M.Y., 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. J. Chiropractic Med. 15 (2), 155–163.

Marra, A.D., Becker, H., Axhausen, K.W., Corman, F., 2019. Developing a passive GPS tracking system to study long-term travel behavior. Transport. Res. C: Emerg. Technol. 104, 348–368.

McHugh, M.L., 2012. Interrater reliability: the kappa statistic. Biochemia Medica 22 (3), 276–282.

Nichols, T.R., Wisner, P.M., Cripe, G., Gulabchand, L., 2010. Putting the kappa statistic to use. Qual. Assurance J. 13 (3-4), 57–61.

Open Science Collaboration, 2015. Estimating the reproducibility of psychological science. Science 349 (6251).

OpenStreetMap contributors, 2020. Planet Dump Retrieved from https://planet.openstreetmap.org, (Data file from July 2020).

R Core Team, 2021. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna; Austria. https://www.r-project.org/.

Ranganathan, P., Pramesh, C.S., Aggarwal, R., 2017. Common pitfalls in statistical analysis: measures of agreement. Perspect. Clin. Res. 8 (4), 187. https://doi.org/10.4103/picr.PICR_123_17.

Rewa, K.C., 2012. An analysis of stated and revealed preference cycling behaviour: a case study of the regional municipality of Waterloo. University of Waterloo.

Safi, H., Mesbah, M., Ferreira, L., 2014. Smartphone-assisted travel surveys: a smart way for transport planning. Sydney, Australia: CAITR (Conference of Australian Institutes of Transport Research).

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

Schüssler, N., Montini, L., Dobler, C., 2011. Improving post-processing routines for GPS observations using prompted-recall data. [Working Paper Transport and Spatial Planning] 724.

Shen, L., Stopher, P.R., 2013. Should we change the rules for trip identification for GPS travel records? Australasian Transport Research Forum, ATRF 2013 - Proceedings.

Shen, L.i., Stopher, P.R., 2014. Review of GPS travel survey and GPS data-processing methods. Transp. Rev. 34 (3), 316–334.

Stopher, P., FitzGerald, C., Zhang, J., 2008. Search for a global positioning system device to measure person travel. Transport. Res. C: Emerg. Technol. 16 (3), 350–369.

Tran, K.A., Barbeau, S.J., Labrador, M.A., 2013. Automatic identification of points of interest in global navigation satellite system data: A spatial temporal approach. In: Proceedings of the 4th ACM SIGSPATIAL International Workshop on Geostreaming, pp. 33–42.

Tsui, S.Y.A., Shalaby, A.S., 2006. Enhanced system for link and mode identification for personal travel surveys based on global positioning systems. Transport. Res. Rec.: J. Transport. Res. Board 1972 (1), 38–45.

Usyukov, V., 2017. Methodology for identifying activities from GPS data streams. Procedia Comput. Sci. 109, 10–17.

van Dijk, J., 2018. Identifying activity-travel points from GPS-data with multiple moving windows. Comput. Environ. Urban Syst. 70, 84–101.

Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. Transport. Res. Record: J. Transport. Res. Board 1768, 125–134.

Xiang, L., Gao, M., Wu, T., 2016. Extracting stops from noisy trajectories: a sequence oriented clustering approach. ISPRS Int. J. Geo-Inf. 5 (3), 29.

Yao, Z., Zhou, J., Jin, P. J., & Yang, F., 2019. Trip End Identification based on Spatial-Temporal Clustering Algorithm using Smartphone GPS Data.

Yeh, S.-C., Hsu, W.-H., Su, M.-Y., Chen, C.-H., Liu, K.-H., 2009. A study on outdoor positioning technology using GPS and WiFi networks. In: 2009 International Conference on Networking, Sensing and Control, pp. 597–601.

Zhou, C., Jia, H., Juan, Z., Fu, X., Xiao, G., 2017. A data-driven method for trip ends identification using large-scale smartphone-based GPS tracking data. IEEE Trans. Intell. Transp. Syst. 18 (8), 2096–2110.