Journal of Transport Geography 100 (2022) 103330

Contents lists available at ScienceDirect

Journal of Transport Geography

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

Assessing the role of geographic context in transportation mode detection from GPS data

Avipsa Roy [a,*], Daniel Fuller [b], Trisalyn Nelson [c], Peter Kedron [d]

[a] Department of Urban Planning and Public Policy, University of California, Irvine, USA
[b] Department of Community Health and Humanities, Memorial University, Canada
[c] Department of Geography, University of California Santa Barbara, USA
[d] School of Geographical Sciences and Urban Planning, Arizona State University, USA

ARTICLE INFO

Keywords:
Supervised learning
GPS
Travel mode detection
Geographic context
Model generalizability

ABSTRACT

The increasing availability of health monitoring devices and smartphones has created an opportunity for researchers to access high-resolution (spatial and temporal) mobility data for understanding travel behavior in cities. Although information from GPS data has been used in several studies to detect transportation modes, there is a research gap in understanding the role of geographic context in transportation mode detection. Integrating the geography in which mobility occurs, provides context clues that may allow models predicting transportation modes to be more generalizable. Our goals are first, to develop a data-driven modeling framework for transportation mode detection using GPS mobility data along with geographic context, and second, to assess how model accuracy and generalizability varies upon adding geographic context. To this extent we extracted features from raw GPS mobility data (speed, altitude, turning angle and net displacement) and integrated context in the form of geographic features to classify active (i.e. walk/bike), public (i.e. bus/train), and private (i.e. car) transportation modes in three different Canadian cities - Montreal, St. Johns, and Vancouver. To assess the role of integrating geographic context in mode detection, we adopted two different modeling approaches – generalized and context-specific, and compared results using random forests, extreme gradient boost, and multilayer perceptron classifiers. Our results indicate that for context-specific models the highest classification accuracy improved by 64% for Montreal, by 74% for St. John's and by 77% for Vancouver compared to the generalized model. We also found that the multilayer perceptron (96%) achieved the highest classification accuracy upon adding contextual variables compared to random forests (94.6%) and extreme gradient boost (93.3%) classifier. Our study highlights that adding contextual information specific to a city's geography can improve the predictive accuracy of transportation mode detection models, however, in case of limited knowledge about the geographic setting of a study area, a generalized model combining GPS data from several cities may still be useful for predicting modes from trip data.

1. Introduction

Understanding the modes of transportation people use to travel within cities is key to planning safer, healthier, and more inclusive environments (Boulange et al., 2017). Detailed information about mobility patterns and transportation mode usage can help planners and policymakers when they make targeted decisions to invest in safe and equitable infrastructure (Nelson et al., 2021; Roy et al., 2019). The growing availability of health monitoring devices and smartphones is now facilitating the collection of high-resolution (spatial and temporal) mobility data, which policymakers can potentially use to overcome methodological problems associated with traditional models (Forrest and Pearson, 2005; Murakami et al., 2004). Such 'big' mobility data provides an opportunity to create a deeper understanding of the transportation mode choices of individuals (Feng and Timmermans, 2013) and to build an aggregate picture of a city's travel behavior (Bohte and Maat, 2009; Chen et al., 2016). Such information can then be used to improve urban infrastructure allocation and enhance accessibility (Ford et al., 2015; Cui et al., 2020) and comfort (Ferster et al., 2021) of a city's residents.

Several studies have been conducted on transportation mode detection either from just GPS data or combining some GIS attributes with the

* Corresponding author at: 300 Social Ecology I, Irvine, CA 92697, USA.
E-mail address: avipsar@uci.edu (A. Roy).

https://doi.org/10.1016/j.jtrangeo.2022.103330
Received 6 July 2021; Received in revised form 9 March 2022; Accepted 19 March 2022
Available online 1 April 2022

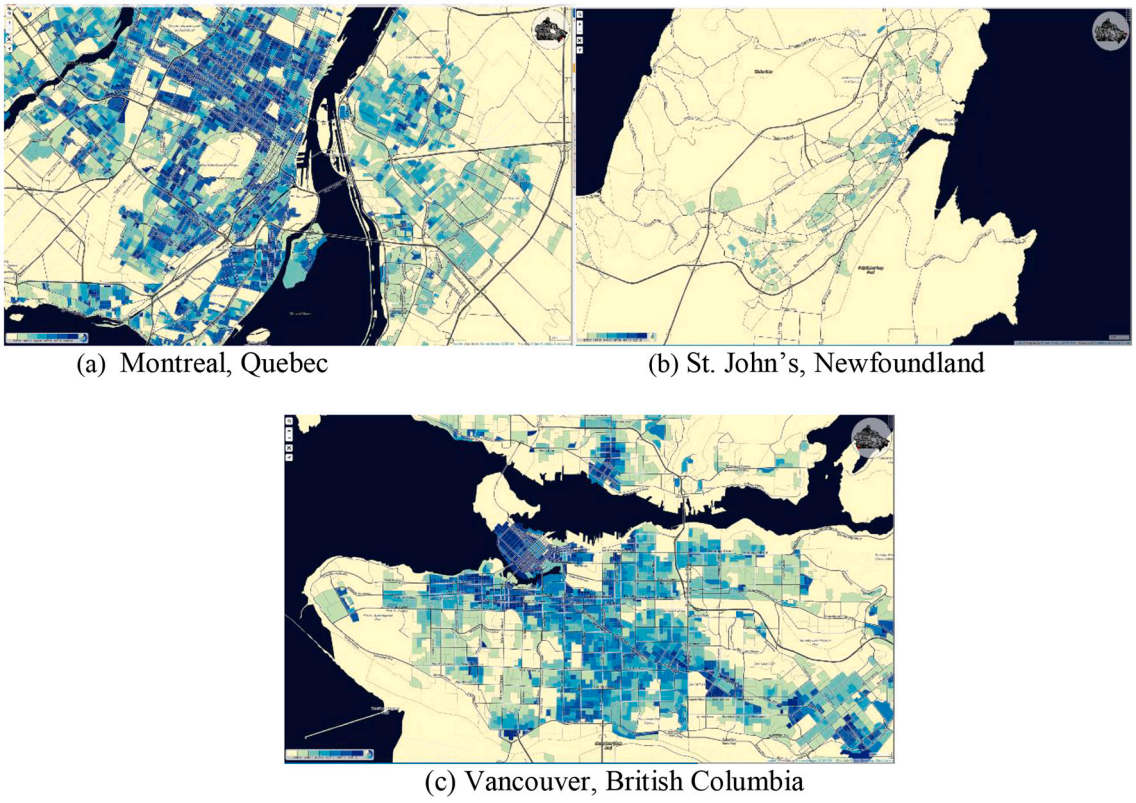

Fig. 1. Maps showing the population density per square miles of three Canadian cities included in the study. (Source: StatCan Census 2016; https://www12.statcan.gc.ca/census-recensement/2021/dp-pd/dv-vd/cpdv-vdpr/index-eng.cfm)

GPS data. Xiao et al. (2015) used four algorithms namely -Support Vector Machines (SVM), Multinomial Logit (MLN), Artificial Neural Networks (ANN), and Bayesian Networks (BN) and found that the BN algorithm performed more accurately than other algorithms (92%). Feng and Timmermans (2013)also used external data and included attributes related to distance from transportation networks, satellite information, and ownership of the vehicle in order to detect the transport mode. Stenneth et al. (2011) detected six transport modes (stationary, walk, bike, car, bus, and train) from raw GPS data by applying the algorithms Random Forests (RF), Naïve Bayes (NB), Multilayer Perceptron (MLP), Decision Trees (DT), and Bayesian Networks (BN). They combined GPS data collected by GPS devices for six individuals over 3 weeks with GIS information, such as transportation networks, bus stations, railway networks, and bus locations. The results showed that the RF algorithm improved accuracy by 17% when they also used GIS information in addition to the GPS data. However, the study by Stenneth et al. (2011) did not consider trip identification for trip segments, and the number of training datasets was very small and comprised only six users. Both Xiao et al. (2015) and Feng and Timmermans (2013) applied attributes, such as average speed, maximum speed, average acceleration, and trip distance. Information from mobility data has also been used in transportation research (Zheng et al., 2008; Auld et al., 2009; Schuessler and Axhausen, 2009; Stenneth et al., 2011; Hemminki et al., 2013) for understanding travel behavior by predicting modes of transportation from GPS and accelerometers (Feng and Timmermans, 2013; Carlson et al., 2015) but without the addition of geographic context.

However, further research is needed to quantify measures of geographic context across multiple cities based upon the city's urban structure and also in assessing how the inclusion of such measures affects the prediction accuracy in comparison to a generalized approach using just GPS features. Currently, it is unclear to what extent the inclusion of contextual measures in the transportation mode detection process could change the classification accuracy. It is therefore essential to understand how geographic factors like the built and natural environment as well as land-use types of a city could influence travel mode choices people make (Wang et al., 2017; Ewing and Cervero, 2010).

While the geographic context may provide clues on modes of travel, geographic data has been used less frequently as a feature for classifying GPS data into travel modes. Traditionally, context has been gathered using data from surveys and questionnaires (Van Vugt et al., 1996, Rodríguez and Joo, 2004, Schwanen and Mokhtarian, 2005, Wener and Evans, 2007). In terms of existing methodologies for mode detection rule-based classifiers have been used in several studies based on a relatively small number of features (Bohte and Maat, 2009; Chen et al., 2010; Gong et al., 2012; Sauerländer-Biebl et al., 2017; Schuessler and Axhausen, 2009; Stopher et al., 2008; Marra et al., 2019). Previous research (Wang et al., 2017; Cheng et al., 2019; Kim et al., 2021) has also shown tree-based approaches like Decision Trees (Shah et al., 2014; Feng and Timmermans, 2013), Random Forests (Wang et al., 2017; Cheng et al., 2019; Nguyen and Armoogum, 2020) and Gradient Boosting (Wang et al., 2018) have proven to be the more appropriate algorithmic approach for mode detection while using GIS information. Although a number of methods exist for classifying transportation modes from GPS data, most of the existing methods are limited in terms of assessing the role of geographic context on predictive accuracy and how they can translate into policies that improve urban life. The inclusion of measures of geographic context in the mode detection process may lead to more accurate predictions needed for effective policy-making, but we have yet to test this hypothesis. However, there is little methodological knowledge on how leveraging varied geographic covariates from multiple sources specific to different cities affects the generalizability of such models.

To address these gaps, we have identified our research goals to examine whether combining GPS and contextual features relevant to several different motorized and non-motorized transport modes from multiple cities can improve the prediction accuracy of mode detection

Table 1
Description of the weather, population, and transportation mode share for each city.[a]

City	Annual monthly temperature ranges		Population	Mode share of commuters			
	Min	Max		Walk	Bike	Public Transit (Bus/Train)	Car
Montreal	−22.3 °C	32.1 °C	4,247,446	5.2%	2.0%	22.3%	8.6%
St. John's	−13.8 °C	27.2 °C	108,860	4.6%	0.2%	3.1%	17.8%
Vancouver	−4.9 °C	26.2 °C	2,463,431	6.7%	2.3%	20.4%	11.2%

[a] Source: Mode share was collected from data provided by Statistics Canada, Commuters using sustainable transportation, https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016029/98-200-x2016029-eng.cfm. The temperature data was provided by Environment Canada from highest and lowest temperatures averaged from 2013 to 2020 https://climate.weather.gc.ca/climate_data/almanac_selection_e.html

and to assess how much the generalizability of the model varies upon adding geographic context to transportation mode detection. To this extent, we first extract meaningful features from the GPS traces and combine these features with contextual information from geographic features guided by existing literature. Then, we train different supervised classification models to predict travel modes in three different Canadian cities and finally validate and assess the generalizability and accuracy of the trained models using just GPS features alone versus combining GPS and contextual features in each city.

We aim to highlight the role of geographic context in improving the prediction accuracy of transportation modes and variation in model accuracy between generalized model and context-specific machine learning models in transportation mode detection. Our study develops methods which are open and reproducible and can be used by practitioners as a guideline to choose appropriate contextual variables for accurately predicting transportation modes as well as testing the generalizability of prediction by combining those variables with new unseen mobility datasets for other cities.

2. Study area

We performed the study across three Canadian cities: Montreal in Quebec, St. John's in Newfoundland and Labrador, and Vancouver in British Columbia (Fig. 1). Each of the cities has emphasized the role of equitable transportation infrastructure in their urban planning policies (City of Vancouver, 2021; St. John's, 2015; Olliere, 2018; Urban Planning and Mobility Department, City of Montréal, 2017) and invested in GPS-based data collection programs aimed at understanding travel mode choice. The study cities also provide the opportunity to study mode choice under variations in population density (Fig. 1), but with comparable spatial and temporal data resolution. The study cities are similarly varied in their weather conditions, population size, and mode share of commuters for different sustainable and active transportation modes (Table 1). This variation in geographic context contributes to the generalizability of our mode detection models.

Briefly, geographic variation in the three study cities can be summarized as follows: Montreal is the cultural and economic hub of the province, with the second largest population in Canada. It is a port city and is surrounded by St. Lawrence and Ottawa rivers. It is a walkable city and is interspersed with bike lanes and bike paths. The city is also well-connected by different public transit modes like subways, buses, and trains connecting the city to the entire province. St. John's is a harbor city with a downtown of steep hills and winding streets. The City of St. John's maintains a road network of over 1400 km, as well as a network of sidewalks for pedestrians and parking infrastructure throughout the city. The Metrobus transit is a popular public transit service in the city and alongside this, the city also maintains a road network of over 1400 km, as well as a network of sidewalks for pedestrians and parking infrastructure throughout the city. Finally, Vancouver boasts an accessible and convenient public transit system with several modes including bus, SkyTrain, ferries as well as bicycles (City of Vancouver, 2018). As the city is surrounded by water on three sides, it has several bridges to the north and south. Although similar to most other cities in that the automobile serves as the primary mode of transportation, it has alternatives such as the SkyTrain system, which is the longest fully automated light metro system in the world, and an extensive network of bicycle paths.

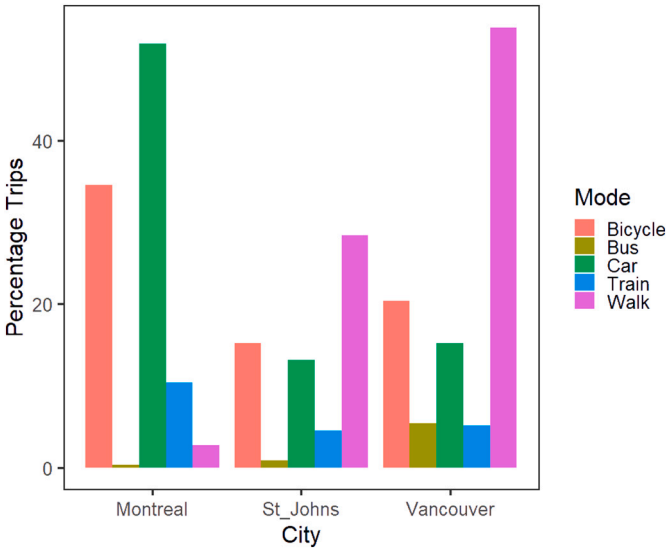

Fig. 2. Trip characteristics collected from GPS devices for multiple cities.

Vancouver being much warmer than the other two cities reflects higher use of active transportation modes including walking and cycling (nearly 10%) (Table 1). Montreal has a well-connected transit network with nearly 22.3% of its commuters using public transit modes as their mode of choice. However, only 7.2% of the population using active modes of transit for commute purposes. St. John's is much smaller in terms of total population and with much harsher climatic conditions which typically lead people to use private vehicles with only about 25.8% people (Table 1) availing active, public or shared transit modes.

3. Data

We used GPS-enabled mobile applications to collect a total 3,226,659 unique user-defined trips from Montreal, St John's, and Vancouver between August and December 2017. Data for St. Johns's and Vancouver were collected through a smartphone application Itinerum (Patterson et al., 2019) which collected GPS data at 1-min temporal resolution. The data for Montreal were collected using the MTL Trajet mobile application (MTL Trajet, 2017; https://donnees.montreal.ca/ville-de-montreal/mtl-trajet) which collected GPS trajectories of user movements from the origin and destinations by truncating to the nearest intersection. The data collection mechanism was similar to that in St. John's and Vancouver as the MTL Trajet uses the same underlying technology as Itinerum devices. All trips were for both St. John's and Vancouver were labeled by participants and for Montreal were inferred by the mobile app using a trip detection algorithm (MTL Trajet, City of Montreal,2020).

The transport modes for each trip were labeled by the respective GPS

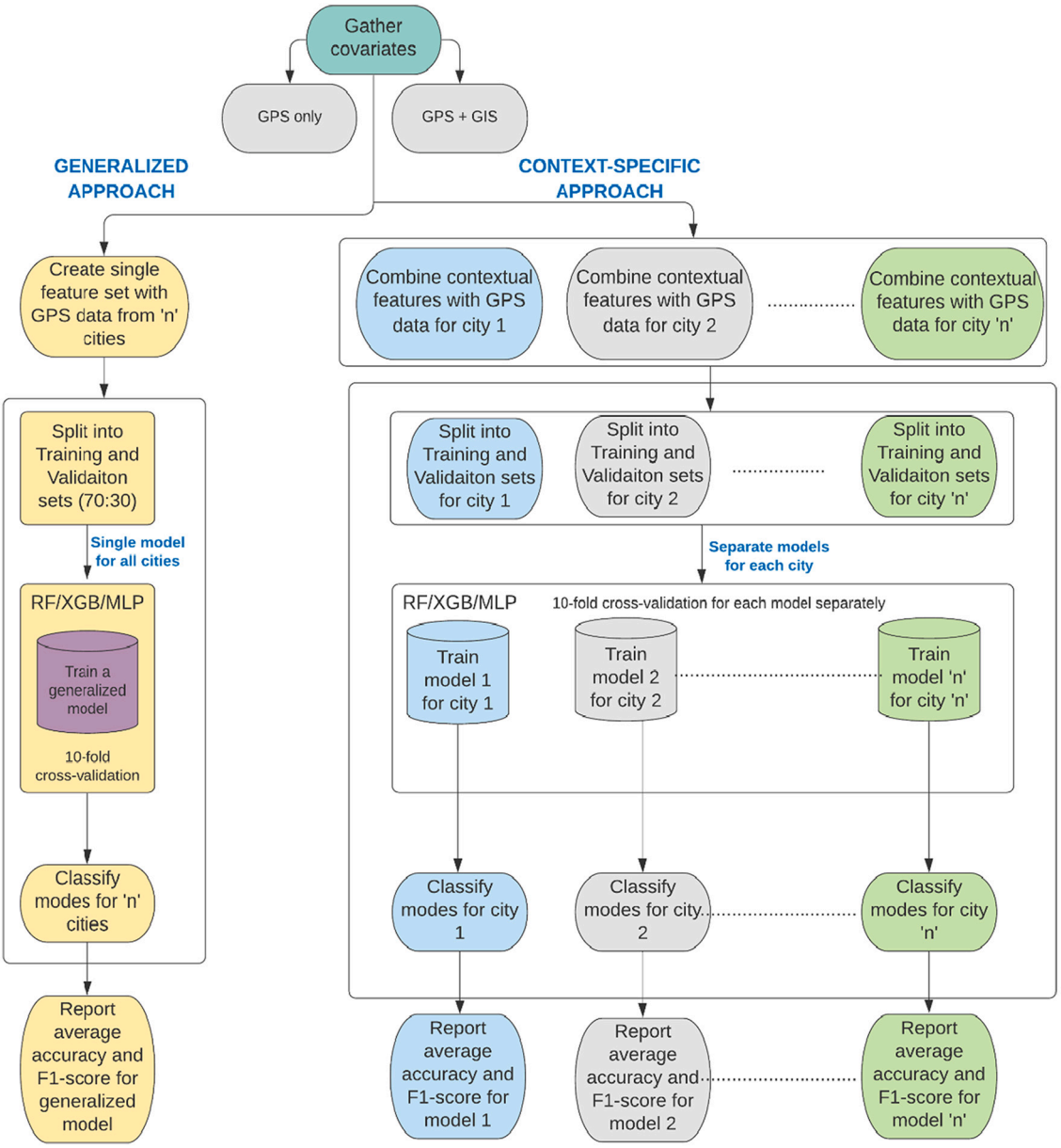

Fig. 3. Overall workflow of transportation mode detection using generalized and city specific approaches.

platforms which captured the trajectories all cities into five different travel from all three cities namely – Bicycle, Bus, Car, Train and Walk (Fig. 2). The percentage of trips falling into each transportation mode are highlighted in Fig. 2. The trips were mostly well distributed between active modes (bicycle, walk) for St. John's and Vancouver, but there were also a significant number (approx. 52%) of motorized vehicle trips recorded in Montreal. The least percentage of trips were available for public transit modes (bus, train) across all three cities. Based upon available data, Vancouver had the highest percentage of walking trips compared to other cities whereas Montreal had the highest percentage of motorized vehicle trips (car, trucks) and bicycle trips were highest for St. John's.

4. Methods

For the prediction of transportation modes we adopted two different modeling approaches – (a) 'generalized' approach using GPS data from all 3 cities to fit a single model and (b) 'context-specific' approach using separate feature sets combining contextual features with GPS data to fit individual models for each city (Fig. 3). We predict five different transportation modes for each approach – 'Bicycle', 'Bus', 'Car', 'Train' and 'Walk'.

We used three classification algorithms for training the generalized and context-specific models- Random Forests (RF), which have been found to have high precision and recall accuracy (Stenneth et al., 2011; Reddy et al., 2010; Ellis et al., 2014; Mäenpää et al., 2017) in classifying motorized and non-motorized transportation modes; Extreme Gradient Boost (XGB), which have also shown considerable success in a wide range of practical applications ((Friedman, 2001); and, Multilayer Perceptrons (MLP), also used in a some earlier studies (Stenneth et al., 2011) to further distinguish between modes from GPS and GIS data.

The feature sets from the GPS and contextual data were constructed using R and RStudio. The generalized and context-specific classification frameworks were modeled in Python using supervised classifiers from the 'scikit-learn' library (Pedregosa et al., 2011). The distance-based proximity metrics used as contextual features were derived using ArcGIS.

Table 2
List of metrics extracted from raw mobility data captured by GPS platforms.

Features	Type	Operationalization	Relevance	References
Speed	GPS	Speed calculated from the consecutive points of the trajectory	Variability in speed can highlight the difference between motorized and non-motorized transport modes.	Stenneth et al. (2011), Zheng et al. (2010), Bohte and Maat (2009), Reddy et al. (2010), Shen and Stopher (2014), Xiao et al. (2015), Roy et al., 2020
Altitude	GPS	The average altitude throughout the trip	The height can indicate whether the user travels in underground subways versus on foot or larger vehicles like buses etc.	Wang et al. (2017), Feng and Timmermans (2013), Roy et al., 2020
Displacement	GPS	The net displacement between consecutive locations along the trajectory	The net displacement can distinguish between motorized and non-motorized transport modes with longer trips taken on public or private modes versus shorter ones are made using active modes.	Zheng et al. (2010), Xiao et al. (2015), Feng and Timmermans (2013), Roy et al., 2020
Turning Angle	GPS	The relative and absolute turning angles of between consecutive points of a trajectory	The orientation can help distinguish a motorized vehicle that can only drive on roads and may not usually turn or change to a new lane unless necessary	Wang et al. (2017), Roy et al., 2020

4.1. Feature extraction from GPS data and geographic context

The GPS records in all datasets were recorded as latitude and longitude and were converted to UTM (Universal Transverse Mercator) coordinates (easting, northing) using pyproj 1.9.5.1. Movement metrics like speed, altitude, net displacement and turning angle (Table 2) were calculated for all GPS records available throughout each city for each participant. The primary unit of analysis are the participants' GPS trajectories over the entire study period converted into trips. Every minute, the GPS device registered the position coordinates (i.e., latitude, longitude, and elevation) of a participant, which were converted into complete trajectories for a single trip using 'R' packages.('adehabitatLT', Calenge and Calenge, 2015).

From the basic movement metrics listed in Table 2 we further applied aggregation functions - mean, sum, standard deviation, skewness, peak intensity, entropy and sum log energy of per trip to extract features combined into a single input matrix which is referred to as 'GPS' features in the following subsections.

Similarly, we refer to the metrics derived from the contextual variables shown in Table 3 as to derive mean, sum, standard deviation, skewness, entropy, peak intensity and sum log energy per user per trip as a separate feature set and call it 'GIS'. To extract contextual information about the surrounding environment through which individuals move, we extracted proximity measures as hausdorff distances based on similarity of trajectories to the shortest path of the nearest points of interest around a GPS trajectory of each user. The points of interest were extracted using a data mining approach from Overpass API and combined with the GPS features using spatial joins. The POIs were then categorized into land-use types such as residential areas, commercial areas, green spaces, and transportation hubs like bus stops, subway stations, bike lanes, and topographic characteristics like distance to the shoreline and comfort level of streets. Attributes like speed (Stenneth et al., 2011; Zheng et al., 2010; Bohte and Maat, 2009; Reddy et al., 2010; Shen and Stopher, 2014; Xiao et al., 2015), acceleration (Stenneth et al., 2011; Roy et al., 2020), proximity to bus stops (Gong et al., 2012; Nguyen and Armoogum, 2020) and rail lines (Stenneth et al., 2011) have been used several times in previous studies, however, proximity to different land-use types and infrastructure specific to active modes of transportation within the context of mode detection are newly introduced in this research.

All features were normalized using a min-max function and used as inputs to supervised classification algorithms mentioned in Section 4.2. Finally, all feature sets were split into training and validation sets using a 70:30 ratio.

Table 3
List of metrics extracted from geographic context using GIS.

Features	Type	Source	Operationalization	Relevance	References
Distance to open space	GIS	OpenStreetMap	Mean hausdorff distance to the nearest open space or green space from the points along the trip trajectory	People using active modes tend to use less traffic-prone areas and closer to open green areas like parks etc.	Roy et al. (2019), Semanjski et al. (2017), Böcker et al. (2015)
Distance to residential areas	GIS	Overapass API	Mean hausdorff distance to the nearest residential area from the points along the trip trajectory	Roads that serve as access to residential areas are used more by bicyclists and pedestrians as they have legal priority over cars, owing to lower speed limits allowing children to play on the street.	Roy et al. (2019), Semanjski et al. (2017),
Distance to commercial centers	GIS	Overapass API	Mean hausdorff distance to the nearest commercial area from the points along the trip trajectory	Typically trips in and around commercial areas are made using public transit modes and cars for office locations, shopping centers. Longer distances may indicate motorized modes are used more compared t bicycling or walking.	Roy et al. (2019), Semanjski et al. (2017)
Distance to subway stations	GIS	OpenStreetMap (OSM, 2017)	Mean hausdorff distance to the nearest subway station from the points along the trip trajectory	Trajectories that are closer to subway stations can be used to identify public transport modes	Gong et al. (2012), Stenneth et al. (2011)
Distance to bus stops	GIS	OpenStreetMap (OSM, 2017)	Mean hausdorff distance to the nearest bus stop from the points along the trip trajectory	Trajectories that are closer to bus stops can be used to identify public transport modes	Gong et al. (2012), Stenneth et al. (2011), Nguyen and Armoogum (2020)
Distance to bike infrastructure	GIS	OpenStreetMap (OSM, 2017)	Mean hausdorff distance to the nearest bike lane/bikeway/bike path from the points along the trip trajectory	Trajectories that are closer to bike infrastructure can be used to identify active transport modes	Roy et al. (2019), Jestico et al. (2016), Semanjski et al. (2017)
Distance to shoreline	GIS	Overpass API	Mean hausdorff distance to the nearest shoreline from the points along the trip trajectory	Trajectories that are closer to the shoreline and have lower speeds can be used to identify active transport modes like biking or walking with an additional aspect of scenic effect for the comfort of pedestrians/ bicyclists. It could also highlight the use of private modes if the trips tend to have higher speed.	Nelson et al. (2021)

Table 4
Model Accuracy comparison between the generalized and city-specific approaches.

Place	Classifier	Mean accuracy		Overall increase in mean accuracy after adding geographic context
		Generalized (GPS features from all cities)	City-Specific (Adding contextual features for individual cities with GPS features)	
Montreal	RF	0.33	0.97	64%
	XGB	0.33	0.89	56%
	MLP	0.60	0.97	37%
St. John's	RF	0.55	0.99	44%
	XGB	0.51	0.98	47%
	MLP	0.25	0.99	74%
Vancouver	RF	0.20	0.88	68%
	XGB	0.20	0.93	73%
	MLP	0.15	0.92	77%

Table 5
Mean classification accuracy metrics for different classifiers using generalized and city specific approaches across all cities.

Classifiers	Generalized model		Context-specific model		Change in accuracy	
	Mean	S.D.	Mean	S.D.	Mean	S.D.
RF	0.36	0.17	0.94	0.05	0.58	0.12
XGB	0.34	0.15	0.93	0.04	0.59	0.13
MLP	0.33	0.23	0.96	0.03	0.62	0.22

4.2. Training generalized and city-specific models to predict transportation modes

The 'generalized' approach was used to fit a single generic model 'G' using the GPS metrics extracted from GPS data irrespective of which city the data was collected from. The combined feature sets from 'n' cities - 'C_1, ….C_n' were then trained using the same model 'G' to predict transportation modes in all cities. Three supervised classifiers namely - random forests (RF), extreme gradient boost (XGB) and multilayer perceptron (MLP) were trained for the 'generalized' approach using data from all cities.

$$G \sim f\left(C_{1,}\, C_{2,} \ldots . C_n\right) \tag{1}$$

The 'context-specific' approach was used to fit one model 'M_i' per city 'i' using all three classifiers. This approach aimed at fitting the model more closely with the geographic characteristics of each city by adding contextual features specific to the city on top of the GPS features represented as 'C_i' for the specific city.

$$M_i \sim f\left(C_i\right) \tag{2}$$

A 10-fold cross-validation was applied for training and testing all the models from Eqs. (1) and (2) for predicting five transportation modes – 'Bicycle', 'Bus', 'Car', 'Train' and 'Walk'. To account for the imbalance in trip distribution across all modes a Synthetic Minority Sampling Technique (Chawla et al., 2002) was applied during the training phase.

4.3. Comparing classification accuracy and assessing generalizability

We computed the F1-score of each model represented by Eqs. (1) and (2) to compare how the classification accuracy varied between the two approaches after adding contextual features. The F1-scores reported are

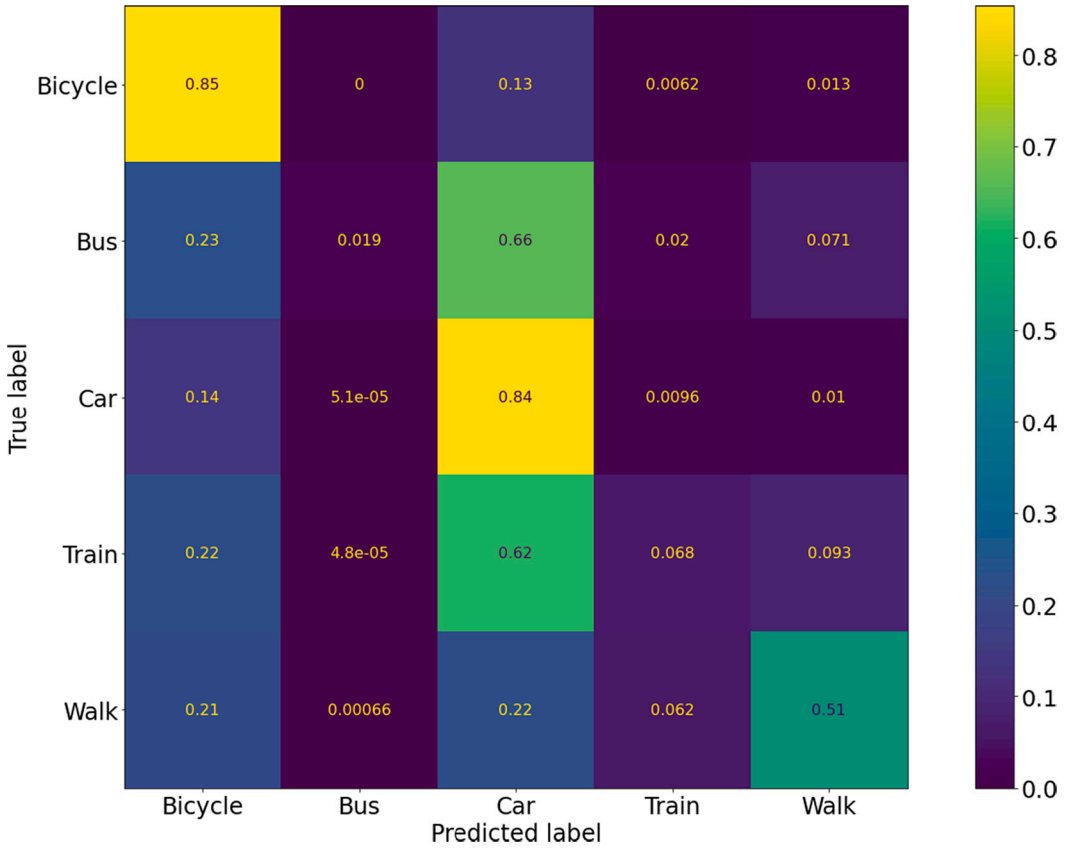

Fig. 4. Confusion matrix for generalized model using GPS features of all cities combined.

calculated based on the precision (Eq. (3)), a measure of the relevance of the results, and recall (Eq. (4)), a measure of how many truly relevant results are returned by the models. A high precision score signifies low false-positive rates, and a high recall indicates low false-negative rates. The F1-score (Eq. (5)) is the harmonic mean of the precision and recall rates which measure the classification accuracy of the models based on true and predicted labels and varies between 0 and 1 with a higher value indicating greater accuracy.

$$P = \frac{T_p}{T_p + F_p} \tag{3}$$

$$R = \frac{T_p}{T_p + F_n} \tag{4}$$

$$F1 = 2^{*}\frac{P^{*}R}{P + R} \tag{5}$$

The F1-scores were also compared between the generalized and city specific models to assess whether a single model or multiple different models would be suitable for prediction purposes when combining metrics from GPS data with surrounding geographic context. The generalized model 'G' used GPS feature sets as input derived features from all cities and the city-specific models 'M_i' used a total of three different feature sets combining GPS and contextual data as inputs- one for each city. The generalized and city-specific models were rerun using the RF, XGB and MLP classifiers.

5. Results

We grouped the trips and mode from GPS variables listed in Table 2 to extract a total of 40 features. The contextual variables were used to generate an additional 68 features for GIS data listed in Table 3. The generalized and city-specific model accuracies with and without geographic context are reported in Table 4. Overall, the results indicated that the accuracy of all three classifiers improved significantly upon adding contextual data.

The overall increase in classification accuracy ranged from 37% up to 77% after GIS features were added to the models along with the GPS features (Table 4). The highest improvement in accuracy was achieved by MLP in case of St. John's and Vancouver as the volume of data was much lower compared to that of Montreal.

We also calculated the summary statistics for different classifiers to

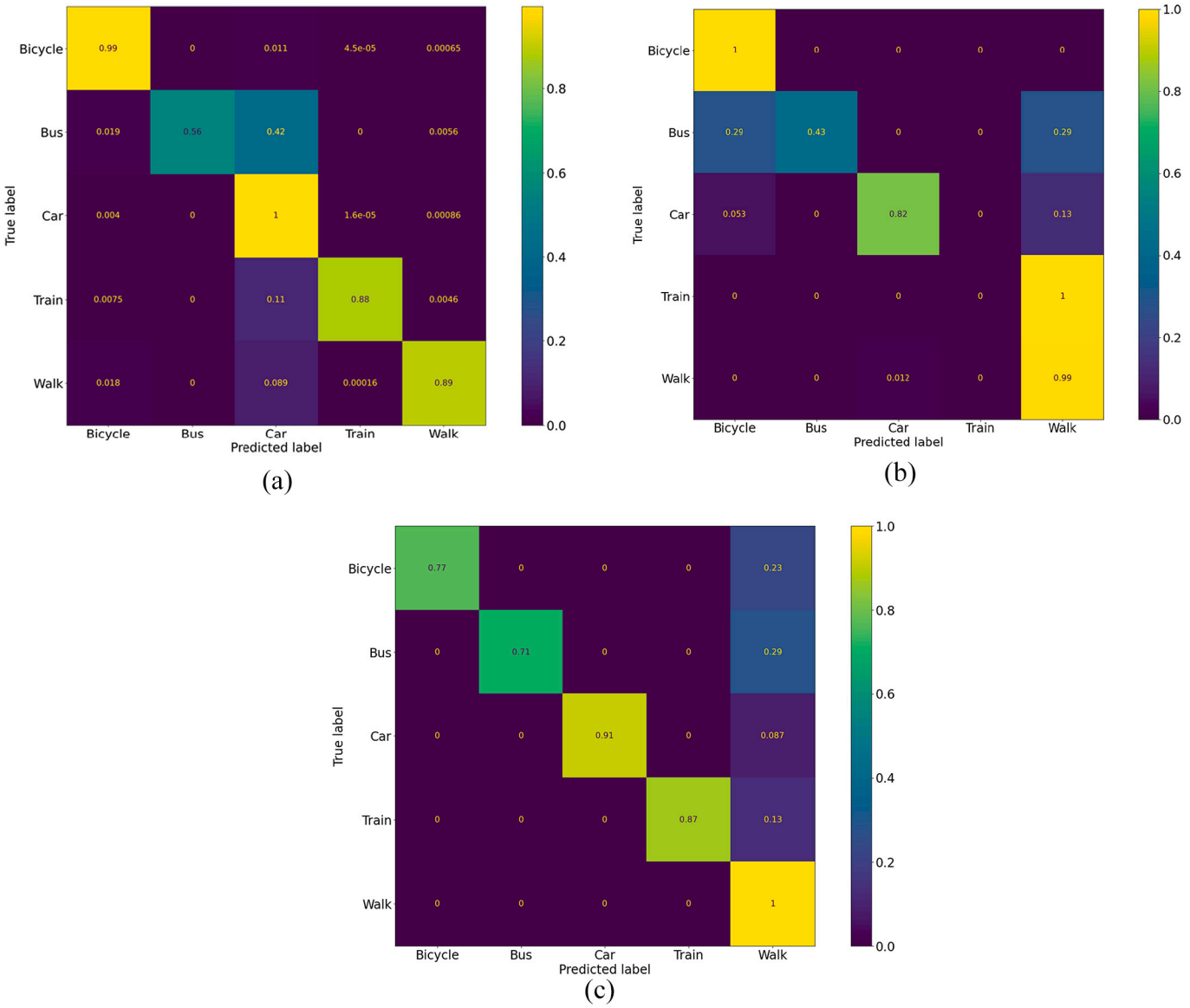

Fig. 5. Confusion matrices for context-specific models after adding geographic context for (a) Montreal, (b) St. John's and (c) Vancouver.

compare predictive accuracy using both approaches and found that the highest average accuracy improved by 62% while using MLP compared to RF which showed 58% improvement and XGB which showed a 59% improvement (Table 4).

The highest average accuracy of 96% was obtained by the context-specific model using the MLP classifier (Table 4). The low standard deviations also indicate that the context-specific approach yielded much more stable models specific to each city, whereas the generalized model not only hurt the classification accuracy which ranged from 33%–36% but also varied greatly based on training data. Additionally, we compared the context-specific models with separate models for each city using just GPS features as shown in Supplementary Tables S1 and S2. The MLP classifier resulted in the most stable model with the lowest variance (Table S2) in prediction accuracy. The accuracy improved significantly upto 68% (Table S1) when just contextual features were added on top of just GPS features. However, the XGB model tends to overfit as the accuracy did not vary a lot.

We used the MLP classifier, as it had resulted in the highest change in accuracy as shown in Table 5, to generate a confusion matrix for trips from all cities combined using the generalized model (Fig. 4). The most misclassification rate occurs in distinguishing between the motorized modes with 66% of car trips being classified as bus trips and 62% of train trips are misclassified as car trips.

However, active modes of transportation have a higher classification accuracy using the generalized approach. Bicycling is classified correctly in 85% of the trips and walking modes are classified correctly in 51% of the trips. Similarly, the city-specific models were also used to generate one confusion matrix per city using the MLP classifier (Fig. 5). The context-specific models performed significantly better than the generalized one. For Montreal with ‘Bicycle’, ‘Train’ and ‘Walk’ were classified with high accuracies of 99%, 88% and 89% respectively with only 56% trips correctly classified as using the mode ‘Bus’. All five modes were correctly classified for St. John's with nearly 99% accuracy for ‘Walk’ modes, whereas in Vancouver ‘Walk’ modes were most accurately classified with other modes ranging from 71% to 91% accuracy. The most misclassification occurred among ‘Bus’ and ‘Train’ trips which were incorrectly predicted as ‘Walk’ trips.

6. Discussion

We developed a data-driven modeling approach to classify transportation modes to assess the role of geographic context in classifying transportation modes across multiple cities. Our results highlight that adding contextual information specific to cities using variables like - distance to bike infrastructure, distance to subways, distance to shoreline and distance to open spaces significantly improves prediction accuracy of transportation modes labeled as ‘Bicycle’, ‘Bus’, ‘Car', ‘Train’ and ‘Walk’. We used distance-based measures to encode proximity to transportation infrastructure (e.g. subway stations, bus stops, and bike lanes) in Montreal, St. John's and Vancouver that were helpful in determining a particular transport mode choice for commuters (Lunke, 2020) as they typically try to optimize their commute times by staying closer to areas with better availability of transportation infrastructure. The natural and built environment of a city (e.g. residential neighborhoods, open spaces etc.) was also an important factor in the transport mode choice as shown by earlier studies (Winters et al., 2010) and contributed to the high accuracy of context-specific models compared to using just GPS metrics using the generalized model (Table 4).

Geographic context captured by variables like distance to commercial areas (Semanjski et al., 2017) and distance to green spaces or seashores (Semanjski et al., 2017; Böcker et al., 2015) are often important for leisure trips made on foot, bicycles or in private vehicles who can spare time to interact more with their spatial surroundings (Páez and Whalen, 2010) as well as feel safe and comfortable (Ferster et al., 2021). Aesthetics like the visibility of the sea-shore (Nelson et al., 2021) or proximity to open green spaces (Roy et al., 2019) that have lower traffic volumes and safer speed limits (Roy et al., 2019) could also be used as a distinguishing factor for active (bicycle/walk) and private (car) modes of transportation as people who ride bicycles tend to prefer more scenic routes versus people who use cars as they tend to prefer shorter routes with higher speed limits. Hence, including these factors considerably improved the accuracy of the classification algorithms for the context-specific approach (Table 5).

The raw GPS data were converted into meaningful features that were combined with the contextual variables to fit supervised classification models and the accuracy of each model improved after adding contextual variables. Our results highlight how considering contextual variables in determining transportation modes can improve overall accuracy. In terms of the generalizability of the models, the changes in accuracy indicate that context-specific models capture the geographic setting of the city quite well building upon how and when people tend to use certain modes of transportation compared to others based on the availability of transit stops, access to bike lanes or other geographic features it is trained with and therefore restricts its applicability in other cities to some extent. The accuracy for Montreal went up by nearly 64% using Random Forests, but it was lower than the other two cities – even though more trip data were available for Montreal. A possible explanation could be that -since trips were labeled using a trip generation algorithm, the uncertainty of the algorithm may have propagated into the classification approach as well leading to lower accuracies than in St. John's or Vancouver.

However, an interesting thing to note is that the accuracy of the context-specific approach sharply increased the classification accuracy upto 99% for some trips in the city of St. John's. A possible explanation could be the data volume of the trips used for fitting the models was not too high and there may have been a chance overfitting of the model. Most of the trip imbalance was accounted for during the training phase, however, since our modeling depends to a large extent upon the availability of real-world data, we were limited in terms of proportionate amount of data for each city. Even though we cross-validated the context-specific model using several combinations of validation data, it may have not been sufficient to overcome the overfitting issue for St. John's as more pre-labeled trips of matching spatial and temporal resolution were not available from the city at the time of this study resulting in a much sparse input feature matrix available for prediction compared to Vancouver and Montreal. Hence, it is advised to consider issues of overfitting carefully while fitting context-specific models and reducing the number of independent attributes used to fit the model ot increasing the volume of groud truth data may help overcome this issue.

Based on the ultimate purpose of classifying trips, practitioners may either choose a highly accurate model or a highly generalizable model. The highly accurate model may produce correct transportation mode labels but would depend on a greater number of available training data and would perform well in a single study area. The highly generalizable model compromises on very high accuracy but will ensure the model will perform optimally well in multiple study areas with varying geographic context and will not be entirely skewed towards any single city or a high amount of readily available correctly labeled trips.

In the future we hope to extend our work to balance out trip distributions from all these cities for comparable amount of ground truth data to further validate our model accuracy for the context-specific approach and a coordinated effort among different local governmental agencies may be needed during the data generation/collection process itself to make the modeling process streamlined. Further context-specific attributes will also be tested beyond the ones currently used in the study (Table 3) to better evaluate the significance of the contextual features on classification accuracy of the context-specific models. Since GPS features are not tied to the local geographic setting (i.e information about the terrain or availability of transportation infrastructure etc.), using the generalized model might be a reasonable and or easier approach to classifying transportation modes as well when policymakers are unaware of which contextual variables may be used for every city. The

predictive capacity of the generalized approach could potentially help in the generation of trip labels from raw GPS data at a large scale to overcome manual effort of labeling trips.

7. Conclusion

Overall our results can inform policymakers to quantify how context influences travel behavior in cities. Our modeling approach is open and reproducible and can be used to predict transportation modes from GPS data and contextual information in other cities depending upon the availability of data. The context-specific methods developed will be applicable in scenarios where the underlying urban structure of the city is to be closely studied and have a significant influence on mode choices. The results generated in this paper could provide a guideline to policymakers on which additional factors to consider for predicting transportation modes beyond the traditional instrumental factors like distance, speed, time, and cost. With further research and refinement of our results policymakers can better understand how and why the travel demand for different transport modes fluctuates with the dynamics of space, time, and place. The results can be utilized in helping them design well-planned data collection efforts for travel behavior studies that could enable more equitable infrastructure investments for one and all. However, the results might vary by geographic settings of a study area and policymakers need to prioritize their goals of higher accuracy versus high generalizability to choose an optimal model that suits their needs.

Acknowledgments

The authors would like to thank INTERACT team for providing valuable feedback and supporting the work. The study is supported by a grant #IP2-1507071C from the Canadian Institutes of Health Research. This study was approved by the Memorial University Interdisciplinary Committee on Ethics in Human Research (20180188-EX).

Appendix A. Supplementary materials

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.jtrangeo.2022.103330.

References

Auld, Joshua, Williams, Chad, Mohammadian, Abolfazl, Nelson, Peter, 2009. An automated GPS-based prompted recall survey with learning algorithms. Transp. Lett. 1 (1), 59–79.

Böcker, Lars, Dijst, Martin, Faber, Jan, Helbich, Marco, 2015. En-route weather and place valuations for different transport mode users. J. Transp. Geogr. 47, 128–138.

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

Boulange, Claire, Gunn, Lucy, Giles-Corti, Billie, Mavoa, Suzanne, Pettit, Chris, Badland, Hannah, 2017. Examining associations between urban design attributes and transport mode choice for walking, cycling, public transport and private motor vehicle trips. J. Transp. Health 6, 155–166.

Calenge, C., Calenge, M.C., 2015. Package 'adehabitat'. R package version 1, 18.

Carlson, Jordan A., Saelens, Brian E., Kerr, Jacqueline, Schipperijn, Jasper, Conway, Terry L., Frank, Lawrence D., Chapman, Jim E., Glanz, Karen, Cain, Kelli L., Sallis, James F., 2015. Association between neighborhood walkability and GPS-measured walking, bicycling and vehicle time in adolescents. Health Place 32, 1–7.

Chawla, Nitesh V., Bowyer, Kevin W., Hall, Lawrence O., Philip Kegelmeyer, W., 2002. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357.

Chen, Cynthia, Gong, Hongmian, Lawson, Catherine, Bialostozky, Evan, 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: lessons learned from the New York City case study. Transp. Res. A Policy Pract. 44 (10), 830–840.

Chen, Cynthia, Ma, Jingtao, Susilo, Yusak, Liu, Yue, Wang, Menglin, 2016. The promises of big data and small data for travel behavior (aka human mobility) analysis. Transp. Res. C Emerg. Technol. 68, 285–299.

Cheng, Long, Chen, Xuewu, De Vos, Jonas, Lai, Xinjun, Witlox, Frank, 2019. Applying a random forest method approach to model travel mode choice behavior. Travel Behav. Soc. 14, 1–10.

City of St. Johns, 2015. Social and equity aspects of … – Memorial University. Retrieved December 17, 2021, from. https://www.mun.ca/harriscentre/aptf2015/Kerry_Murray_pres.pdf.

City of Vancouver, 2018. Transportation Panel Survey 2018. https://vancouver.ca/files/cov/2018-transportation-panel-survey.pdf.

City of Vancouver, 2021. Social Equity & Regional Growth Study - Metro Vancouver. Social Equity and Regional Growth. Retrieved December 17, 2021, from. http://www.metrovancouver.org/services/regional-planning/PlanningPublications/MVSocialEquity-RegionalGrowthStudy.pdf.

Cui, Boer, Boisjoly, Geneviève, Miranda-Moreno, Luis, El-Geneidy, Ahmed, 2020. Accessibility matters: exploring the determinants of public transport mode share across income groups in Canadian cities. Transp. Res. Part D: Transp. Environ. 80, 102276.

Ellis, Katherine, Godbole, Suneeta, Marshall, Simon, Lanckriet, Gert, Staudenmayer, John, Kerr, Jacqueline, 2014. Identifying active travel behaviors in challenging environments using GPS, accelerometers, and machine learning algorithms. Front. Public Health 2, 36.

Ewing, Reid, Cervero, Robert, 2010. Travel and the built environment: a meta-analysis. J. Am. Plan. Assoc. 76 (3), 265–294.

Feng, T., Timmermans, H.J., 2013. Transportation mode recognition using GPS and accelerometer data. Transp. Res. C Emerg. Techn. 37, 118–130.

Ferster, Colin, Nelson, Trisalyn, Laberee, Karen, Winters, Meghan, 2021. Mapping bicycling exposure and safety risk using Strava Metro. Appl. Geogr. 127, 102388.

Ford, Alistair C., Barr, Stuart L., Dawson, Richard J., James, Philip, 2015. Transport accessibility analysis using GIS: assessing sustainable transport in London. ISPRS Int. J. Geo Inf. 4 (1), 124–149.

Forrest, Timothy L., Pearson, David F., 2005. Comparison of trip determination methods in household travel surveys enhanced by a global positioning system. Transp. Res. Rec. 1917 (1), 63–71.

Friedman, Jerome H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 1189–1232.

Gong, Hongmian, Chen, Cynthia, Bialostozky, Evan, Lawson, Catherine T., 2012. A GPS/GIS method for travel mode detection in New York City. Comput. Environ. Urban. Syst. 36 (2), 131–139.

Hemminki, Samuli, Nurmi, Petteri, Tarkoma, Sasu, 2013. Accelerometer-based transportation mode detection on smartphones. In: Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems, pp. 1–14.

Jestico, Ben, Nelson, Trisalyn, Winters, Meghan, 2016. Mapping ridership using crowdsourced cycling data. J. Transp. Geogr. 52, 90–97.

Kim, Kyusik, Kwon, Kyusang, Horner, Mark W., 2021. Examining the effects of the built environment on travel model choice across different age groups in Seoul using a random forest method. Transp. Res. Rec. 03611981211000750.

Lunke, Erik Bjørnson, 2020. Commuters' satisfaction with public transport. J. Transp. Health 16, 100842.

Mäenpää, Heikki, Lobov, Andrei, Martinez, Jose L., Lastra., 2017. Travel mode estimation for multi-modal journey planner. Transp. Res. C Emerg. Technol. 82, 273–289.

Marra, Alessio D., Becker, Henrik, Axhausen, Kay W., Corman, Francesco, 2019. Developing a passive GPS tracking system to study long-term travel behavior. Transp. Res. C Emerg. Technol. 104, 348–368.

Murakami, Elaine, Wagner, David P., Neumeister, David M., 2004. Using global positioning systems and personal digital assistants for personal travel surveys in the United States. In: International Conference on Transport Survey Quality and Innovation.

Nelson, Trisalyn, Roy, Avipsa, Ferster, Colin, Fischer, Jaimy, Brum-Bastos, Vanessa, Laberee, Karen, Hanchen, Yu, Winters, Meghan, 2021. Generalized model for mapping bicycle ridership with crowdsourced data. Transp. Res. C Emerg. Technol. 125, 102981.

Nguyen, Minh Hieu, Armoogum, Jimmy, 2020. Hierarchical process of travel mode imputation from GPS data in a motorcycle-dependent area. Travel Behav. Soc. 21, 109–120.

Olliere, M., 2018, August. At the crossroads of Sustainable Transportation and social … At the Crossroads of Sustainable Transportation and Social Inclusion. Retrieved December 17, 2021, from. https://www.socialconnectedness.org/wp-content/uploads/2019/10/At-The-Crossroads-of-Sustainable-Transportation-and-Social-Inclusion-2.pdf.

OpenStreetMap contributors, 2017. Planet Dump. Retrieved from. https://planet.osm.org. https://www.openstreetmap.org.

Páez, Antonio, Whalen, Kate, 2010. Enjoyment of commute: a comparison of different transportation modes. Transp. Res. A Policy Pract. 44 (7), 537–549.

Patterson, Zachary, Fitzsimmons, Kyle, Jackson, Stewart, Mukai, Takeshi, 2019. Itinerum: the open smartphone travel survey platform. SoftwareX 10, 100230.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, et al., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Reddy, Sasank, Mun, Min, Burke, Jeff, Estrin, Deborah, Hansen, Mark, Srivastava, Mani, 2010. Using mobile phones to determine transportation modes. ACM Trans. Sensor Networks (TOSN) 6 (2), 1–27.

Rodríguez, Daniel A., Joo, Joonwon, 2004. The relationship between non-motorized mode choice and the local physical environment. Transp. Res. Part D: Transp. Environ. 9 (2), 151–173.

Roy, A., Fuller, D., Stanley, K., Nelson, T., 2020. Classifying transport mode from global positioning systems and accelerometer data: a machine learning approach. Findings 14520.

Roy, Avipsa, Nelson, Trisalyn A., Stewart Fotheringham, A., Winters, Meghan, 2019. Correcting bias in crowdsourced data to map bicycle ridership of all bicyclists. Urban Sci. 3 (2), 62.
Sauerländer-Biebl, Anke, Brockfeld, Elmar, Suske, David, Melde, Eric, 2017. Evaluation of a transport mode detection using fuzzy rules. Transp. Res. Procedia 25, 591–602.
Schuessler, Nadine, Axhausen, Kay W., 2009. Processing raw data from global positioning systems without additional information. Transp. Res. Rec. 2105 (1), 28–36.
Schwanen, Tim, Mokhtarian, Patricia L., 2005. What affects commute mode choice: neighborhood physical structure or preferences toward neighborhoods? J. Transp. Geogr. 13 (1), 83–99.
Semanjski, Ivana, Gautama, Sidharta, Ahas, Rein, Witlox, Frank, 2017. Spatial context mining approach for transport mode recognition from mobile sensed big data. Comput. Environ. Urban. Syst. 66, 38–52.
Shah, Rahul C., Wan, Chieh-yih, Hong, Lu, Nachman, Lama, 2014. Classifying the mode of transportation on mobile phones using GIS information. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 225–229.
Shen, Li, Stopher, Peter R., 2014. Review of GPS travel survey and GPS data-processing methods. Transp. Rev. 34 (3), 316–334.
Stenneth, Leon, Wolfson, Ouri, Yu, Philip S., Bo, Xu., 2011. Transportation mode detection using mobile phones and GIS information. In: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 54–63.
Stopher, Peter, FitzGerald, Camden, Zhang, Jun, 2008. Search for a global positioning system device to measure person travel. Transp. Res. C Emerg. Technol. 16 (3), 350–369.
Urban Planning and Mobility Department, City of Montréal, September 05, 2017. Se déplacer... et y gagner. Retrieved April 05, 2021, from. https://ville.montreal.qc.ca/mtltrajet/.
Vugt, Van, Mark, Paul A.M., Lange, Van, Meertens, Ree M., 1996. Commuting by car or public transportation? A social dilemma analysis of travel mode judgements. Eur. J. Soc. Psychol. 26 (3), 373–395.
Wang, Bao, Gao, Linjie, Juan, Zhicai, 2017. Travel mode detection using GPS data and socioeconomic attributes based on a random forest classifier. IEEE Trans. Intell. Transp. Syst. 19 (5), 1547–1558.
Wang, Bijun, Wang, Yulong, Qin, Kun, Xia, Qizhi, 2018. Detecting transportation modes based on LightGBM classifier from GPS trajectory data. In: 2018 26th International Conference on Geoinformatics. IEEE, pp. 1–7.
Wener, Richard E., Evans, Gary W., 2007. A morning stroll: levels of physical activity in car and mass transit commuting. Environ. Behav. 39 (1), 62–74.
Winters, Meghan, Brauer, Michael, Setton, Eleanor M., Teschke, Kay, 2010. Built environment influences on healthy transportation choices: bicycling versus driving. J. Urban Health 87 (6), 969–993.
Xiao, G., Juan, Z., Zhang, C., 2015. Travel mode detection based on GPS track data and Bayesian networks. Comput. Environ. Urban. Syst. 54, 14–22, 14e22.
Zheng, Yu, Li, Quannan, Chen, Yukun, Xie, Xing, Ma, Wei-Ying, 2008. Understanding mobility based on GPS data. In: Proceedings of the 10th International Conference on Ubiquitous Computing, pp. 312–321.
Zheng, Yu, Chen, Yukun, Li, Quannan, Xie, Xing, Ma, Wei-Ying, 2010. Understanding transportation modes based on GPS data for web applications. ACM Trans. Web (TWEB) 4 (1), 1–36.