Travel Behaviour and Society 39 (2025) 100936

Contents lists available at ScienceDirect

Travel Behaviour and Society

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

Studying transfers in informal transport networks using volunteered GPS data

Genevivie Ankunda *, Christo Venter

Department of Civil Engineering and Centre for Transport Development, University of Pretoria, Private Bag X20, Pretoria, South Africa

ARTICLE INFO

Keywords:
Machine learning
Transfers
Walking time
Waiting time
Walking distance
GPS

ABSTRACT

Multimodal integration is an important issue in public transport systems due to its influence on both passenger experience and overall network efficiency. In most countries in the global South, achieving integration is particularly problematic because of the informal nature of most public transport. Decentralised service planning and demand responsiveness lead to often uncoordinated, highly variable service patterns, which are not optimised from a passenger perspective. Efforts to promote integration are also hampered by a lack of planning data on routes, service frequencies, and transfer locations. This research asks whether GPS data supplied by passengers as they move through the network can be used to help form a better understanding of the extent and quality of the transfer experience. The data was collected in the City of Tshwane, South Africa, among informal minibus-taxi passengers. Post-processing involved the use of a machine learning algorithm to identify in-vehicle, wait and walk segments, which were used to identify transfers between one vehicle and another. The results showed that many transfers are spatially efficient with short walk and wait times, but that a minority of transferring passengers may experience very long transfers. Transfers encompass a diverse range of behaviours including pacing, shopping and browsing, and typically involve much more walking than waiting. Transfers also occur across a wide range of locations, but tend to be concentrated in certain nodes and along street segments. Strategies to improve transfer facilities as well as general walkability might be targeted at such locations. The study demonstrated that volunteered GPS data is a promising source of information to help planners understand the transfer experience in multimodal networks in data-poor environments.

1. Introduction

Multimodal integration is an important issue in public transport systems. Transport integration is the process of bringing the elements that comprise transport systems into closer and more efficient interaction across modes and operators, to improve the overall state and quality of services (NEA et al., 2003). Many city and regional authorities have explicit goals and policies to pursue integration, recognising that it has benefits both to passengers (in terms of improved coverage and reduced travel disutility), and to operators (through improved network efficiency) (Kash and Hidalgo, 2014, Aziz et al., 2018, Ceder, 2021).

In most countries in the global South, achieving integration is problematic because of the informal nature of most public transport. Informal services, also termed paratransit, artisanal, or popular transportation (Behrens et al., 2015), are characterised by fragmented ownership, the use of small vehicles, high degrees of demand responsiveness, and operations largely falling outside the ambit of government planning and regulation (Kumar et al., 2021). Although these services take many forms, they collectively provide the majority of urban mobility services throughout Africa, Latin America, and most of developing Asia (Behrens et al., 2021). The decentralised service planning of informal services leads to often uncoordinated, highly variable service patterns. Routes are typically established by operators (individually or in groups) (Kerzhner, 2023), and service areas delineated to maximise profit and to manage competition amongst rival operators rather than to optimise passenger convenience or connectivity (Cervero and Golub, 2007). Some evidence suggests that informal networks often impose fewer transfers but higher transfer penalties on passengers than formally planned bus and rail services (Tun et al., 2020). However, the extent and characteristics of transfers in informal networks have been understudied, hampering the ability of authorities to promote integration and improve passenger service quality.

A key limitation in this regard is the lack of planning data of all types, including on routes, service frequencies, and transfer locations. Informal

* Corresponding author.
E-mail addresses: ankunda.genevivie@gmail.com, u21742040@tuks.co.za (G. Ankunda), christo.venter@up.ac.za (C. Venter).

https://doi.org/10.1016/j.tbs.2024.100936
Received 2 July 2024; Received in revised form 17 October 2024; Accepted 23 October 2024
Available online 12 November 2024

services suffer from this in particular; Klopp and Cavoli (2017) call them "invisible" within formal urban planning processes. Emerging data collection and data analytics techniques offer potentially promising ways to help fill the data gap. One example is Global Positioning System (GPS) technology, a location awareness system which relies on a network of satellites to determine the ground position of an object (Stopher, 2009a). The high level of location accuracy of GPS, coupled with its relatively easy deployment using smartphones, offers ways of collecting travel information with greater ease and at lower cost than what can be done through traditional survey and observational techniques (Stopher, 2009a). GPS has proved useful to track paratransit vehicles in order to map routes and service patterns (Klopp and Cavoli, 2017, Saddier and Johnson, 2018, Coetzee et al., 2018). Although transfer locations can be inferred from GPS data processed in the General Transit Feed Specification (GTFS) format, such data are less useful in studying network-wide integration, as tracking occurs at the vehicle and route level, and contains limited information on the actual locations where transfers happen as passengers are navigating through the network.

An alternative use of GPS technology has been to track passengers instead of vehicles. For instance, crowdsourcing, the process of collecting and utilising the creative solutions of a distributed network of individuals (Goodchild, 2007, Howe, 2008), has been used to collect data on bus routes and generate the first bus map of Dhaka (Ching et al., 2013). This is an example of Volunteered Geographic Information (VGI), a type of crowdsourced information which is linked to geolocation data or a map (Ferster et al., 2018). VGI has also been used to study aspects of the urban travel experience more generally (Howe, 2021).

This research asks whether tracking data that are supplied by passengers, as they move through an informal public transport network, can be used to help form a high-level understanding of the transfer experience. Key elements of this experience, that are often perceived negatively by the public, are walking, and waiting, and transferring (Fang and Zimmerman, 2015). In order to identify and measure these elements, we develop and test a method to collect self-reported VGI data using a smartphone application. The resultant datasets are characterised by large volumes of complex data, requiring advanced computer-based processing methods for efficient management and analysis. To help identify data patterns associated with transferring, we use machine learning, a field of artificial intelligence which involves programming computers to optimise a performance criterion, based on past experience or a sample of training data (Alpaydin, 2020). The outcomes are potentially instructive as a method for researchers and city authorities to study the transfer experience in multimodal networks in data-poor environments.

The specific research questions are:

Is it feasible to use data collected through a smartphone-based GPS application, supplied by passengers in a crowdsourced setting, and a machine learning approach to data analysis, to identify the physical locations of transfers in informal transport networks; and Can the amount of walking time, waiting time, and walking distance be estimated from the GPS data, as a way of quantifying the passenger disutility of transferring?

This research was conducted in the City of Tshwane, a sprawling metropolitan area in South Africa. Informal services are provided with 16-seater minibus-taxis (MBT), a mode that transports about two-thirds of public transport trips (and referred to as *taxis* in the rest of the paper). Formal bus and rail services also operate, but fall outside the scope of this study. The rest of this paper is organised into the following sections: literature review, materials and methods, results, and conclusions.

2. Literature

This section reviews literature on informal public transport networks, the measurement of the quality of transport integration, the use of GPS-based data collection, and machine learning applications in transport research.

2.1. Informal public transport service patterns

Following the rising recognition that informal public transport services in global South cities require better study (Cervero and Golub, 2007, Behrens et al., 2015), research examining their service patterns, operating conditions, and passenger impacts is slowly expanding. Much of the research is grounded in qualitative case studies and small samples, but a few larger studies using GPS data have also appeared (Molloy et al., 2023, Yazdizadeh et al., 2019, Costa et al., 2023, Fan et al., 2019, Du Preez et al., 2019, Mittal et al., 2024, Ndibatya and Booysen, 2021).

Within Sub-Saharan Africa (SSA), a common finding has been that informal networks tend to provide very good coverage of a metropolitan area. A comparative study of seven SSA cities found that in most cities nearly three-quarters of residents live within a 10-minute walk of mapped paratransit routes, making the mode very widely available to potential users (Falchetta et al., 2021).

However, this coverage is not universal. Service patterns are typically very variable, both across space and time, in response to variations in demand density, traffic conditions, and the presence of competitor modes (Ferro, 2015). Spatially, many informal routes tend to extend between outlying residential areas and central cities (Falchetta et al., 2021), following major desire lines for radial movements. Two implications are that routes tend to become very dense in the city centre, contributing to major congestion in already crowded areas; and that passengers travelling between suburbs are often forced to make long journeys with inconvenient transfers. Du Preez et al. (2019) found that in Cape Town this is mitigated by the existence of two other types of paratransit services, which they labelled intermediate and feeder/distributor services. These tend to serve medium and short trip lengths respectively, with the latter serving a collector role in the vicinity of transfer hubs where passengers can transfer to trunk services provided by both other taxis and formal bus and rail modes. In general routing efficiency tends to be high (Mittal et al., 2024), although some drivers are observed to engage in detours, random searching behaviour, and trips abandoned before the end of the route, which impose delays and uncertainty on passengers (Ferro, 2015, Ndibatya and Booysen, 2020, Ndibatya and Booysen, 2021). This is a consequence of the fact that routes and stops are typically selected by drivers under conditions of high uncertainty, and not optimised from a network operational perspective.

Temporal variations are driven chiefly by the profit motive, as many vehicles are rested during low-demand periods of the day. Long headways are often coupled with 'fill-and-go' dispatch strategies, often leading to long waits for boarding passengers (McCormick et al., 2015). Many vehicles spend the majority of the day waiting or queuing at ranks or stations, leading to low overall utilisation (Saddier and Johnson, 2018, Ndibatya and Booysen, 2020).

2.2. Assessing the quality of multimodal integration

Multimodal integration of public transport networks has been studied using a variety of approaches. A key aspect of most studies is their focus on transfers – points of intersection between public transport lines within the network, where users have to or choose to move from one vehicle to another (Garcia-Martinez et al., 2018).

Transfers are regarded negatively because they disrupt the travel experience, reduce the competitiveness of public transport compared to the door-to-door service provided by private transport, deter potential customers, and reduce the satisfaction of existing customers. On the other hand, transfers support hierarchical multimodal networks and increase the service areas of public transport systems (Guo and Wilson, 2011).

Transfers have predominantly been studied using two methods: experience assessment, and supply assessment. Experience assessment

characterises the transfer experience into the walking and waiting undertaken by commuters during transfers, and a subjective transfer penalty or disutility determined by the transfer environment (Garcia-Martinez et al., 2018, Schakenbos et al., 2016). The underlying aim is to understand transfer behaviour generally, how it is subjectively valued by users, and how it may be improved (Guo and Wilson, 2011). Stated choice techniques are commonly used for collecting data from users and modelling this behaviour. Supply assessment, on the other hand, registers and ranks all elements of transfer supply, including station design, social environment, and service management, based on input from a wide range of stakeholders.

A diverse set of indicators have been used by different researchers to evaluate multimodal integration, the most common being transfer time which typically includes transfer walking time and transfer waiting time (Guo and Wilson, 2011). Some studies have added to this in-vehicle time, smoothness of transfer, and availability of information (Chowdhury et al., 2014, Ceder, 2007, Ceder et al., 2009). The widely used Bus Rapid Transit (BRT) Planning Guide considers walking distance, number and size of transfer points, the need for exit and re-entry into the station, fare payment method, and presentation of information to assess the quality of integration of BRT with other public transport modes (Institute for Transportation and Development Policy (ITDP), 2016). A further dimension in the evaluation of multimodal integration that has gained research attention is the complementarity of service spans between scheduled and unscheduled services. This includes exploring the interventions which can be implemented at transfer points to ensure a seamless and high-quality service for passengers (Plano et al., 2020, Plano and Behrens, 2022).

Some studies have sought to combine these quantitative and qualitative indicators into a coherent framework to express the quality of integration using a single measure. An example is the framework developed by Chowdhury et al. (2014) to measure public transport network connectivity. Rodrigue (2024) defines transport network connectivity as the extent to which passengers travel from one location to another, through a direct connection or through an indirect series of nodes within a transportation network. The level of inter-route and inter-modal connectivity is expressed as a weighted sum of measurable indicators across routes and paths in a network, that might include quantitative indicators such as ride time, walking time, and waiting time, and qualitative indicators such as smoothness of transfer and availability of information.

Moodley and Venter (2022) adopted a similar approach in Durban, South Africa, to develop a multimodal integration index based on individuals' importance ratings for the various dimensions of transferring. The study showed that passengers perceive a range of dimensions, some of which are more easily measurable (such as transfer times and integrated ticketing), and some of which are more subjective (such as personal security, universal access, and traffic safety). Elements relating to comfort and convenience such as shelter, seating, ablutions, overcrowding, and short walking distances were the most important for participants at all the sites surveyed. Notably, this study also recommended the development of a mobile application to automate the measurement of waiting times and walking distances rather than depending on the assessments of researchers during in-field audits.

2.3. Travel data collection

The emerging complexity of urban transportation systems has motivated the evolution of traditional transport data collection methods to suit dynamic transportation environments. In the initial stages of the field of urban transport planning, interviews were the principal method of obtaining information about people's movement (Stopher, 2009b, Hayduk, 1997). However, these methods of self-reporting are limited by factors such as high costs, labour intensiveness, long data collection periods and cycles, transcription errors, low and declining response rates, and trip misreporting due to imperfect memory (Stopher, 2009b, Ehrlich et al., 2020).

Global Positioning System (GPS) technology has emerged as a major source of data able to address some of these limitations (Stopher, 2009a). The high level of location accuracy of GPS, coupled with the emergence of smartphones with GPS sensors, has spurred much research into smartphone-based GPS applications for mobility data collection (Stopher, 2009a, Bricka et al., 2014). Some of the applications include the Route Choice Application (RAPP-UP) (Hayes and Venter, 2022), GoMetro (Coetzee et al., 2018, Ndibatya et al., 2017), Flocktracker (Ching et al., 2013, Palencia Arreola, 2019, Yun et al., 2019), TransitWand (Klopp et al., 2015), and Sparrows (Joseph et al., 2020).

Mobility data such as trip origin, destination, travel time, routes used, and number of trips, are captured relatively accurately by GPS technologies (Krygsman and Nel, 2009). Additionally, smartphone-based data collection methods are argued to reduce survey costs, increase capability for larger sample sizes, extended data collection periods and shorter data collection cycles, while offering unprecedented detail of route choice and opportunities for gathering feedback from survey participants (Asakura et al., 2014, Ching et al., 2013). Owing to the multiple sensors contained in smartphones, these devices can generate significant amounts of data, leading to the need for more sophisticated big data analysis techniques (Coetzee et al., 2018, Goenaga et al., 2023).

Digital tools have also been used to systematically collect, analyse, and share Volunteered Geographical Information (VGI) through crowdsourcing. Some characteristics of VGI include high accessibility and shareability; collective data management; and collaborative data collection from contributors through implicit or explicit observations (Yan et al., 2020). These characteristics shape crowdsourcing as a valuable tool for responding easily to local needs and providing more spatial and temporal coverage in data collection compared to traditional approaches (Ferster et al., 2018). On the other hand, VGI is also characterised by a general lack of data quality standards; privacy and security concerns around its legal collection, use and dissemination (Yan et al., 2020); and incomplete representation due to sampling and response bias (Brown, 2017) which all constitute the main barriers to adoption for decision making (Ferster et al., 2018). Nevertheless, VGI seems useful for exploratory research applications.

The data volunteering method of data collection has been applied in some transport and spatial planning studies in South Africa (Howe, 2021), and globally (Ching et al., 2013). Ching et al. (2013) experimented with guided crowdsourcing, also known as flock sourcing, to produce the first bus map of Dhaka including mapping and directions data, travel time, wait time, speed, crowding, service disruptions, safety conditions, and user perception. Flock sourcing is a form of crowdsourcing where users are organised and motivated to set targets, participate, and be held accountable in collection of data for a specific use, instead of relying on a pool of online volunteers as in traditional crowdsourcing. The more systematic nature of flock sourcing can be used to overcome the limitations of traditional crowdsourcing data collection endeavours and was therefore adopted for this research.

2.4. Data post processing and machine learning in transport data analysis

GPS data post-processing typically involves mobile and network computing activities such as data cleaning, stay and move identification, and map matching.

2.4.1. Data cleaning

Data cleaning is the process of removing errors from GPS location positioning data to enable extraction of the trip sequence (Asakura et al., 2014). Some causes of errors in GPS tracking include the urban canyon effect which causes loss of signal precision, poor positioning of the GPS logger, insufficient number of visible satellites at any given time during the tracking process, or complete signal loss due to obstructions such as tunnels and tall buildings (Auld and Mohammadian, 2014).

2.4.2. Stay and move identification

Stay and move identification is a density-based, spatial clustering algorithm used in activity-based modelling to identify stop and move points from 3-dimensional GPS data based on a threshold value of the distance between consecutive points (Gong et al., 2015). The main application of this technique is to split a complete trip from origin to destination into segments whose ends are activities (stops). It has also been applied to evaluate the level of performance of transport systems, and determine attributes of travel behaviour such as origin, destination, and departure and arrival times (Asakura et al., 2003).

2.4.3. Map matching

Map matching is the assignment of the observed "move" points onto the most likely link of the transport network (Asakura et al., 2014). Map matching is necessary because of errors which occur in GPS measurements due to road network complexities, inadequate GPS data collection procedures or a combination of these issues causing the locations observed to be located off the actual road centreline (Blazquez and Miranda, 2015).

2.4.4. Mode detection

Mode detection can also be done alongside activity detection (Feng and Timmermans, 2014). Mode detection is the process of segmenting a multi-modal trip into trip segments of a single mode (Kohla et al., 2014), and identifying the mode. The two-step procedure followed by the mode detection model used by Kohla et al. (2014) included identification of the trip sections travelled with predictable modes such as walking, and classification of the other segments as one of the seven modes specified by the model.

Many automatic mode detection approaches have been applied in travel behavior analysis, employing different machine learning algorithms such as the random forest algorithms, support vector machine (SVM), deep or shallow neural networks, among others (Zheng et al., 2010, Bolbol et al., 2012, Kohla et al., 2014, Dabiri and Heaslip, 2018, Fang et al., 2017).

2.4.5. Machine learning

Machine learning is a field of artificial intelligence which involves programming computers to optimise a performance criterion, based on past experience or a sample of training data. Machine learning can be broadly categorised into supervised learning and unsupervised learning (Alpaydin, 2020). Unsupervised machine learning is the process by which models aim to learn the underlying patterns of the input data, without prior access to any output values (Lison, 2015). Supervised machine learning is the process of predicting a target variable based on a model which has been trained and calibrated on a dataset containing both input variables and target variable labels (Janiesch et al., 2021, Lison, 2015). Supervised machine learning for classification purposes involves a finite set of classes which are predetermined before the learning process, and the classification algorithms are used to map the input space into the predefined classes (Nasteski, 2017).

In addition to the application of machine learning in automatic mode detection, conventional machine learning techniques such as the random forest method, support vector machine (SVM) and deep or shallow neural networks have been used widely in transport research such as traffic state prediction (traffic speed and flow), travel time prediction, bus arrival time prediction (Varghese et al., 2020), modelling energy consumption of electric buses (Basso et al., 2023), accident detection (Aiswarya et al., 2023), and image detection (recognition of traffic signals, vehicles or pedestrians in different applications) (Ciregan et al., 2012, Ouyang and Wang, 2013).

In this research, a supervised random forest classification model (Breiman, 2001) was used to predict the classes of the different segments of a trip as either walking, waiting, or in-vehicle movement. The random forest algorithm is a non-linear model that involves aggregating or averaging the predictions of a series of randomised decision trees, each created using a sub sample of the data (Biau and Scornet, 2016). The averaging nature of this model helps to improve the classification accuracy by maximising outcomes from all trees for classification, rather than depending on the outcome from one decision tree. This classification technique is versatile and accurate in large-scale and multi-dimensional feature applications, easily adapts to ad hoc learning tasks, returns information of predictor variable importance (Fernández-Delgado et al., 2014, Biau and Scornet, 2016), and it has a reduced likelihood of overfitting (Aiswarya et al., 2023). A more in-depth description of the machine learning algorithm and its application is outside the scope of this paper, but the interested reader can refer to Breiman (2001), Biau and Scornet (2016) and Parmar et al. (2018).

3. Materials and methods

3.1. Case study description

The City of Tshwane is the administrative capital of South Africa and part of the urban province Gauteng. The city has a population of approximately 3.65 million people, spread out across a sprawling metropolitan area of about 6300 km^2 (Fig. 1) (City of Tshwane, 2023).

As a result of segregative apartheid spatial planning practices, land use in Tshwane is characterised by a spatial divide between the high-density, low-income residential areas located on the fringes of the city and the low-density, higher-income residential developments closer to the city centre. The low-income settlements include both formal and informal housing, and contain largely marginalised communities dependent on public transport to travel long distances to access job opportunities (City of Tshwane, 2015, McKay et al., 2017).

Historically, the city developed as a monocentric city but over time, with changing investment patterns, increasing car ownership, and development of several other satellite nodes, has evolved into a polycentric, multi-nodal urban form (City of Tshwane, 2013). Nevertheless, most jobs are located in or near the urban core. As a result, most public transport operates radially. The modal distribution of all trips is 33 % private car, 29 % walking, 22 % minibus-taxi, 11 % bus, and 3 % rail (City of Tshwane, 2015). Minibus taxis are considered an informal mode, with routes and schedules largely determined by drivers and their associations, and city authorities applying a "light-touch" regulation to control vehicle safety and route permits. Three main types of services are provided by minibuses: local feeder services within residential areas; line-haul services serving longer intra-urban trips between residential areas and the CBD; and inter-urban services connecting to nearby cities and towns. It is not known what proportion of taxi passengers have to transfer during their trip, but of those who transfer, the majority are taxi-taxi transfers, followed by a smaller share of taxi-train transfers as the train network also offers line-haul services to the CBD (Gotz et al., 2015).

Integration between various public transport services is an explicit policy goal of the City of Tshwane (City of Tshwane, 2015). Efforts to promote integration include closer cooperation between the municipality and the taxi industry, especially during the roll-out of the City's nascent Bus Rapid Transit network (Mokoma and Venter, 2023). However Manana (2021) reports that there is currently no formal monitoring, evaluation and feedback process to improve modal integration in the City of Tshwane, partly due to the prohibitive cost of data collection (Department of Transport, 2016). The method described in this paper might help to fill this data gap.

3.2. GPS tracking app

As a first step, the proposed methodology of this research required identification of a method to collect GPS data using wearable devices. An off-the-shelf GPS tracking application called GeoTracker was tested in-house to determine whether quantitative indicators such as waiting time and walking time can be extracted from GPS data. The proof-of

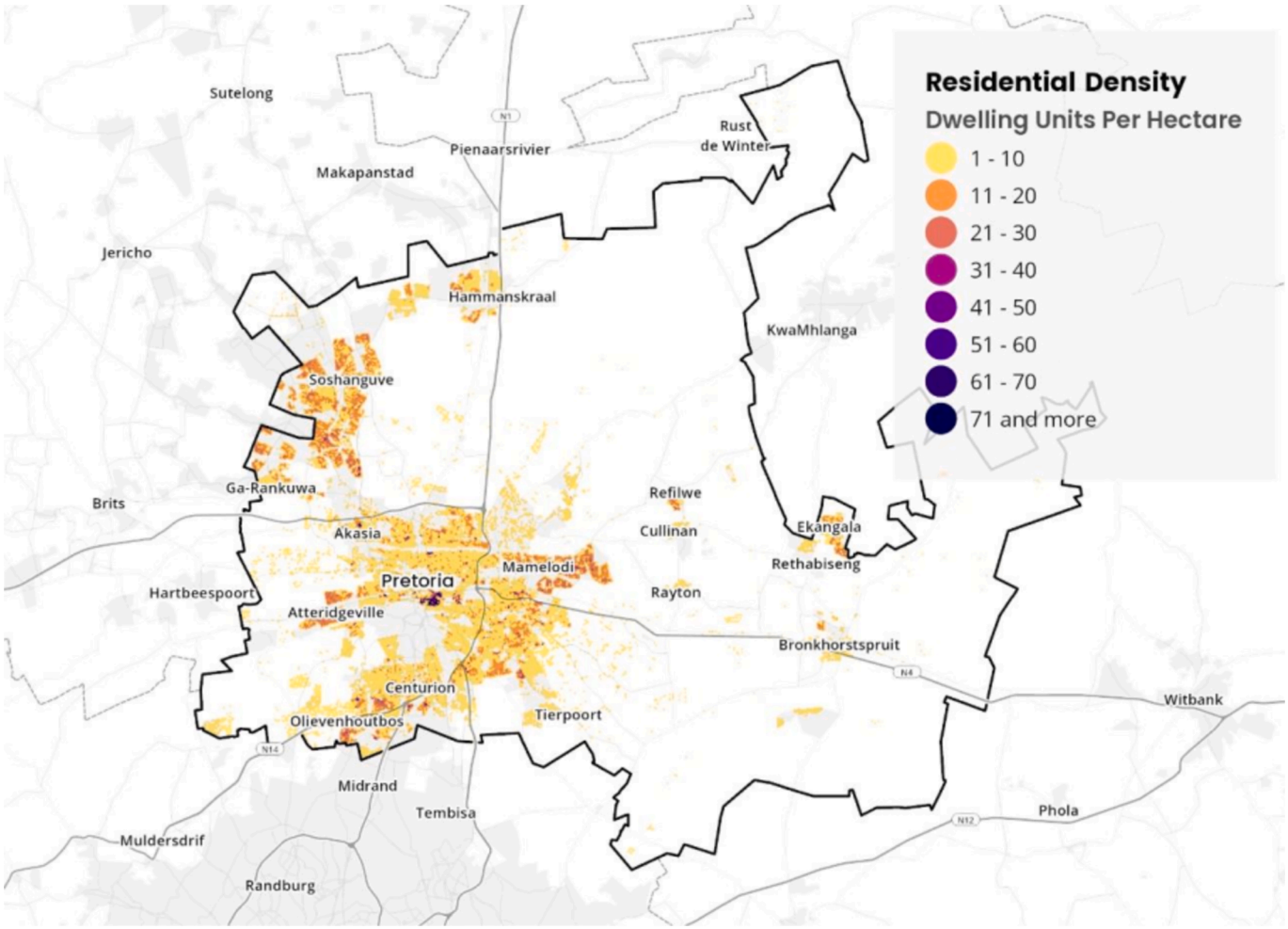

Fig. 1. Residential density of the City of Tshwane (City of Tshwane, 2023).

concept study suggested that it was technically feasible to extract such data, but that the data collection process had to be improved in two ways. Firstly, the app had to provide better granularity and frequency of data outputs to enable efficient data processing; and secondly, the app needed to transmit data in real time to a remote database to mitigate against the risk of data loss due to loss of the host device, and to minimise the respondent burden of manually uploading trip data to the database.

In order to overcome the limitations of the proof-of-concept study, a bespoke GPS tracking smartphone application named TraceMate was developed to incorporate the necessary features. The app ran only on Android smartphones (for now) and had a very simple interface that required a respondent only to press a record button at the start and end of every trip. The app had to be kept running during the entire trip.

3.3. Field work

The study involved collection of quantitative data in the form of GPS traces, and qualitative data (including users' narrative descriptions of their trip sequences, and subjective ratings of different elements of the transfer experience) for each of the trips made through an online questionnaire. The data collection and analysis process is illustrated in Fig. 2.

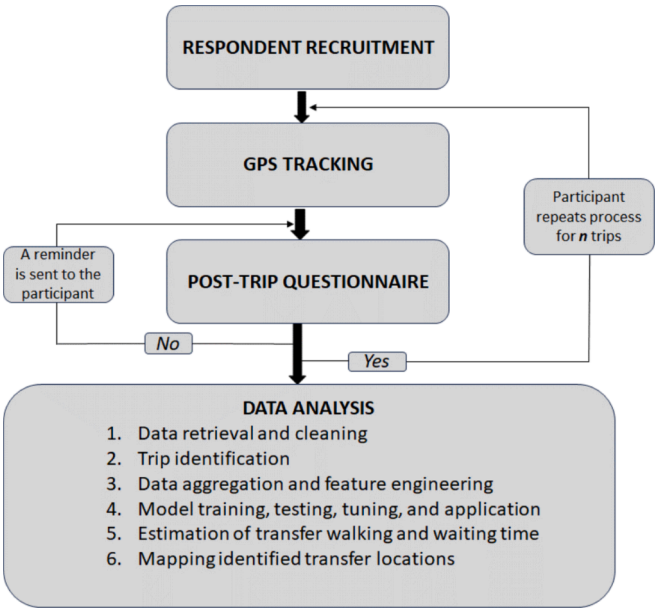

Fig. 2. Data collection and analysis process.

3.4. Sampling and recruitment

As a feasibility assessment, this study did not aim to recruit a

statistically representative sample of commuters. A purposive sampling technique was used based on geographical location to ensure a variety of exemplar trips across the city were observed.

The initial sample included 21 volunteers (i.e., real commuters who recorded their usual trips to and from work or school), recruited via a community-based research organisation. The volunteers were incentivised by reimbursing the costs of mobile data and transport fares used for all the trips recorded, with no extra remuneration. However, as was found in other studies (Howe, 2021), recruiting volunteers is a difficult and time-consuming process. It was therefore decided to supplement the sample with eight paid data collection staff (referred to as enumerators). They were given pre-specified destinations to travel to, as opposed to the random trips of volunteers. The specified destinations were selected after identifying the top ten origin–destination pairs amongst minibus-taxi users in Tshwane (from prior surveys), and selecting those which most likely involved taxi-taxi transfers. Care was also taken to recruit enumerators who were regular public transport users, with knowledge of how to use the systems.

The criteria for recruiting both volunteers and enumerators were: use of at least two modes or two vehicles of the same public transportation mode on their recorded trips; ownership of an Android smartphone; possession of a Google account or willingness to create a temporary account for this survey; informed consent given to have their trips tracked in accordance with ethics and privacy concerns; and basic English proficiency to complete the questionnaires.

3.5. Pilot and main survey

A pilot study was carried out before the main field data collection exercise, with the objectives to act as a pre-test of the newly developed TraceMate GPS tracking application, and to test the practicality of all the proposed steps of data collection. The sample for the pilot study consisted of 12 people recruited from the main sample. Data collection for the pilot survey was done over one day, and debriefing offered the following general observations with regards to the survey tools and the proposed methodology:

Mobile phone battery drainage revealed the need for backup power supply to avoid tracking disruptions. This was critical for participants recording very long trips, several trip repetitions per day (enumerators) and even volunteers whose phone batteries may be drained by the end of the workday.

Information seeking by enumerators who were travelling unfamiliar routes increased their transfer time dramatically compared to ordinary commuters who were already familiar with the location of interchanges, stations, or the trip fares. This implies that the total transfer time estimated from only smartphone data requires additional contextual information such as the route familiarity of the user.

Some participants also reported long traffic-related stops such as mechanical breakdown of vehicles, or interruptions by the traffic police, which can mistakenly be identified as transfers during data processing.

There was also a limitation of the app's device compatibility, even among Android smartphones, for brands which no longer support Google services.

Based on these findings, enumerators were provided with power banks to provide a backup source of power.

Each participant tracked their trips over a period of four days including both weekdays and weekends outside of national holidays in order to reflect the typical traffic patterns. The volunteers recorded 2 trips per day, whereas enumerators recorded 6 trips per day, in the morning, afternoon and evening peak periods. After every trip, participants were required to complete an online questionnaire to provide some qualitative insights about the transfers made during the just completed trip.

The GPS data collected was transferred from the users' devices through real-time transmission from the app directly to a remote database hosted on MongoDB.

3.6. Sample characteristics

The sample of 29 individuals consisted of more women (64 % of participants) than men, consistent with women's higher use of minibus-taxis (Statistics South Africa, 2022). The largest proportion of the sample were aged between 18 and 30 years, followed by the age group from 31 to 40 years. Whereas most respondents reported being unemployed, other participants were undertaking full-time employment, part-time employment, full-time study or part-time study as their main occupation. The sample is not meant to be statistically representative of taxi users but is consistent with the general characteristics of the taxi user population.

Since it was a requirement for participants to own a smartphone, the sample was likely skewed in favour of more advanced users of mobile phones. Since volunteered geographic data depends on smart phone ownership, it is a potential limitation of the research that it may fail to reach all segments of the travelling public. However some studies indicate that smartphone penetration in South Africa had exceeded 90 % by 2019 (ICASA, 2020). Our survey showed that respondents routinely use mobile data to access the internet, suggesting that the smartphone is the primary method for accessing the web (Fig. 3).

3.7. Data for training the classification algorithm

In supervised machine learning, the algorithm learns the patterns of the different classes based on a dataset where the classes are already labelled.

The data for this purpose was collected by the researcher ahead of the main data collection process. This was done by travelling and recording 8 multistage trips using public transport. The times during which transfers, waiting, and walking took place were recorded on a detailed travel diary, and used to label the corresponding records in the GPS file as either WALK, WAIT, or RIDE segments. The class WALK refers to the activity of walking between the drop-off point of the vehicle from which one is transferring, to the boarding location for the next vehicle. The class WAIT was collectively used to indicate (1) waiting for a vehicle at a bus station or taxi rank; (2) in-vehicle waiting for other passengers to board or during the scheduled dwell time for the case of scheduled services such as the train and buses; and (3) waiting which involves walking/pacing around stations or stops. The class RIDE refers to the time spent in a moving vehicle.

The sample size of the training dataset is relatively small. However it proved adequate for the purposes of developing a prototype model to test the feasibility of the proposed approach before committing resources to collecting a larger dataset for model training, as the labelled datasets for training supervised machine learning models involve

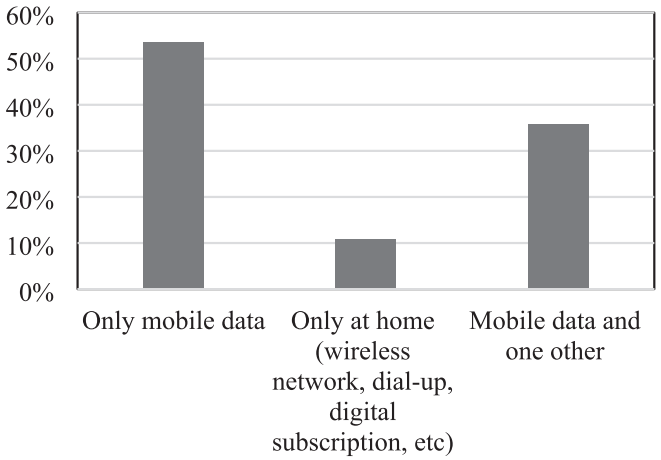

Fig. 3. Routine internet access pathways on mobile phones of sample (n = 29).

manual human annotation which is time-consuming and error prone (Alzubaidi et al., 2023). The small training dataset was highly effective for two reasons: firstly, it was very accurate since it was collected by the researcher; secondly, the ensemble learning and averaging approach used by random forest models reduces sensitivity to noise in small datasets, and helps to maintain a good model quality even in cases where the individual trees are trained on small datasets (Pedregosa et al., 2011, Parmar et al., 2018, Han et al., 2021).

3.8. Data analysis

The data analysis process involved the identification of trips, feature engineering and extraction, training and testing of the model, and its application for prediction of trip segments on trip data supplied by volunteers.

3.9. Data retrieval and cleaning

The collected GPS data was downloaded from MongoDB in form of text files with the variables of object identifier, user identifier, latitude, longitude, instantaneous speed, and a datetime stamp. Initial data processing and cleaning was done in Microsoft Excel, and the cleaned data was imported into R Studio for detailed analysis.

3.10. Trip identification

The first analysis step was to split the cleaned GPS data into distinct trips. We define a 'trip' here as a movement from origin to destination, which could include multiple trip segments with transfers between vehicles. Automatic identification of trip ends was done using the dwell time method (Yang et al., 2022) which was implemented using the 'dplyr' and 'lubridate' packages in R. The dwell time method uses the temporal characteristics of GPS trajectories to identify trip ends, with stop time thresholds serving as a key parameter. Various time thresholds were tested to identify trip ends, leading to a threshold of 600 s (10 min) being selected as the most suitable. This is longer than the 120 s thresholds used in previous studies (Venter et al., 2014, Jianchuan et al., 2014) that was found too short to avoid misclassification of data loss incidents as trip ends where incidents lasted longer than 120 s. A major cause of such incidents was signal loss due to power outages at cellular towers, which was common in Tshwane at the time. In addition, since we mainly tracked the morning and evening commutes, there were significant time intervals between the observed trips, making a longer threshold acceptable.

The output of the trip identification process was validated by comparing the results for each user with their questionnaire responses, which included self-reported detail on all trip start and end times.

3.11. Data aggregation and feature engineering

As a result of several factors related to GPS satellite signals, GPS receivers, and the usage environment (Thin et al., 2016), GPS data inherently contains unavoidable scatter that may lead to inaccurate variations in speed and bearing calculated on a point-to-point basis. To reduce this inaccuracy, trip data was aggregated into segments of 5 s, using a sliding (or rolling) window with an overlap of 4 steps.

From the aggregated data, different features could be extracted. A feature is a property derived from the raw input data with the purpose of providing a suitable representation that is more meaningful to the machine learning process (Janiesch et al., 2021). The most common features applied in mode detection models are speed-based attributes (e.g. average speed, maximum speed, 1st quartile of speed, and 3rd quartile of speed), acceleration-based attributes (e.g. average, minimum, and maximum acceleration), attributes based on the direction of travel (e.g. change of azimuth and the heading change rate), and spatial attributes such as the road or rail network for fixed-route modes (Jianchuan et al., 2014, Ferrer and Ruiz, 2014). ANOVA testing for variable selection in automatic mode detection modelling by Bolbol et al. (2012) concluded that speed and acceleration are the most suitable parameters for differentiating modes, despite the high positive correlation between speed and acceleration that may introduce bias in a mode detection model.

For this analysis, a variety of features calculated within a rolling 5-second window were tested. The most effective were found to be the average, 3rd quartile, and variance of speed, followed by the distance covered within a time window (based on the speed and timestamps), and the change in direction between the start and end points of a time window.

3.12. Variable importance testing

Variable importance testing is necessary for identification of redundant, noisy or unreliable variables which may impair the performance of the final prediction of the classification algorithm. This is important in justifying the cost of data gathering and storage, improving the understandability of the model, and optimisation of computational speed by using only the truly important variables (Han et al., 2016).

In decision trees, the variable importance can be estimated using the Gini Impurity or Gini Index. The Gini Index is a measure of the node impurity in a decision tree. It measures how well a node splits the dataset between the number of outcomes at each node, based on the given conditions. The higher the value of the mean decrease Gini (Gini importance) score, the higher the importance of the variable in the model (Martinez-Taboada and Redondo, 2020).

3.13. Model training and testing

In this analysis, the target variable of the model is the trip segment, a categorical variable with the classes of WALK, WAIT, and RIDE. The predictor variables are the features listed above.

A random forest classification model was trained using the labelled data collected by the researcher and split into 'train' and 'test' subsets in a ratio of 80:20. The 'train' subset was used by the algorithm to learn the patterns of the different categories of the target variable, and thereafter the 'test' subset, which had previously not been exposed to the model, was used to assess the performance of the trained model. A confusion matrix was used to assess the performance of the classification algorithm. A confusion matrix provides a summary of the prediction done on the test dataset by indicating the true positives, false positives, true negatives, and false negatives.

The hyperparameters of the random forest model were tuned to improve the accuracy of prediction. The hyperparameters for a random forest classifier are the number of decision trees (ntree) and the number of predictor variables to randomly sample for consideration at each nodal split in each tree (mtry).

3.14. Model application and estimation of transfer metrics

The trained and tested random forest classifier algorithm was then used to classify the trips from the larger sample, which had been identified by the trip identification algorithm, into segments of walking, waiting, or in-vehicle travel. The trip segments predicted by the random forest classifier for each trip were then summarised by run length encoding, a form of lossless data compression which summarises sequences with the same value occurring many consecutive times into a single value of the count of the particular category (Techie Delight, 2023).

Results of the run-length encoding process were exported from R into Microsoft Excel and used to estimate transfer waiting and walking characteristics. The geographic coordinates for trip origins, destinations, transfer locations, and additional information such as total trip time, could also be extracted for each unique trip.

The data points which were classified as either walking or waiting were mapped in ArcGIS Pro to identify the locations where transfers are made, in addition to trip starts and ends.

4. Results and discussion

4.1. Feature extraction, model training and testing

A summary of the trips tracked for this research, based on the post-trip questionnaire responses received, is shown in Table 1. Out of the planned 240 trips, a total of 230 questionnaire responses covering 205.48 h of travel time were collected.

On the other hand, the automatic trip identification process, which used a threshold of 10 min to identify trip ends, identified a total of 249 trips and an overall total travel time of 155.95 h. The precision of the trip end identification algorithm, defined as the ratio of derived true trip ends to the number of actual trip ends (Reinau et al., 2014), was estimated as 0.92. The shortfall in precision can be explained by instances where i) a single trip had GPS tracking gaps for durations longer than 10 min, and was therefore identified as two separate trips; ii) enumerators ended one trip, and started recording another trip before 10 min had elapsed, causing the algorithm to identify the two trips as a single trip; or iii) respondents failed to self-report all trips on the questionnaire. The 24 % discrepancy between the total travel time recorded by the GPS tracker and self-reported travel times is a concern and is most likely attributable to inaccuracies in the self-reported trip start and end times. Such errors are likely to lie in the extremes of the trip length distribution, as the average duration of reported trips of 53 min (Table 1) is close to the mean travel time of 50 min measured in other surveys in Tshwane (City of Tshwane, 2015).

The relative importance of each extracted feature in predicting the class of the trip segment, based on the Gini importance, is shown in Fig. 4.

From the results shown, the 3rd quartile of speed was the most important feature for predicting the class of the trip segment, followed by average speed, and variance of speed in each data segment. The change in bearing had the lowest Gini importance and was therefore the least influential predictor variable for the model. The random forest classifier model was trained on the training dataset. Using a random search, the model was tuned and the optimal values of ntree and mtry were found to be 250 and 2 respectively.

After training the model, it was then validated on the 'test' subset of the dataset, and an accuracy of 98.84 % was achieved (Table 2). This accuracy was considered satisfactory, and therefore the model was saved and applied for classification of the trip segments on the GPS data collected from commuters.

4.2. Model application

The trained random forest model was then applied to the unlabelled GPS data collected by the survey participants. The predictions from the model were then summarised into sequential WALK, WAIT, and RIDE segments and their respective object lengths (time in seconds) by run length encoding, exported to Microsoft Excel, and visualised using horizontally stacked bar graphs as illustrated in Fig. 5.

The details of transfer waiting and walking for each trip were then estimated from the trip visualisations as well as the run-length-encoded data. A few anomalies were picked up, in the form of multiple short segments identified as WALK segments of less than 1 min. When plotting the coordinates in GIS it was found that the algorithm tended to misidentify low-speed vehicle movements such as at the approach to traffic signals, as walk trips. This was corrected by applying a threshold of 60 s as the minimum segment length for a walk or wait segment, in line with the 60-second minimum cycle length used for traffic signals in South Africa. The example in Fig. 5 (b) shows that a transfer occurred between cumulative object lengths 1700 and 2000, while the short WALK events outside of this window are ignored. Walking at the start and end of trips were also removed from the database, in order to focus on transfer events alone. It was also observed that some trips did not involve any recognisable transfers, as shown in Fig. 5 (a). These trips were omitted from the rest of the analysis.

The total walk and wait time associated with each transfer was calculated. The walking distance was estimated from the walking time, by assuming an average walking speed of 1 m/s as suggested by Hitge and Vanderschuren (2015). This approach was chosen over summing the straight-line distance across individual GPS datapoints because the latter had not been corrected for GPS measurement errors.

Fig. 6 shows the cumulative distribution of the observed walking times for all identified transfer events. The walking times ranged from 1.0 to 26.2 min, with an average walk of 5.13 min, or 308 m. The distribution of walk times is clearly left-skewed, with 88 % of transfer walk times less than 10 min or 600 m. This suggests that transfers tend to be relatively efficient spatially, rarely requiring very long walks between vehicles. However, there is a small minority of transfers that impose much longer walking distances on passengers.

The waiting times observed by time of day are shown in Fig. 7. The waiting time of the sampled trips ranged from 0 to 5.1 min, with an average waiting time of 0.19 min. The figure shows that the majority of transfers had zero waiting time. This is interesting. Firstly, very short waiting times are consistent with the high frequencies of paratransit services, with up to 325 departures per hour on some routes within Tshwane (De Beer and Venter, 2024). It could also indicate that "pure" waiting, where passengers stay in one place waiting for a taxi, is relatively rare for this mode, and that much of the transfer experience is actually made up of walking between egress and access points rather than waiting. This is consistent with the typical stopping pattern of taxis, as shown below. We also observe a negative correlation between walking and waiting, with the longest waiting times estimated for the trips with shortest walking distances. But it is also acknowledged that the distinction between waiting and walking is sensitive to the classification approach adopted, which needs to be refined further in subsequent work.

To reflect the effect of taxis' 'fill-and-go' operation mechanism on the quality of transfers, we plotted both waiting and total transfer time by time of day in Fig. 7. Consistent with the estimated walking and waiting time, 86 % of all trips had transfer times of less than 10 min. However, the majority of trips with transfer times exceeding 10 min occurred during off-peak periods (between 10:00 and 15:00, and after 18:00), with the maximum transfer time observed being 26 min. This suggests that lower frequencies and 'fill-and-go' operations do contribute to somewhat longer transfers for passengers.

Table 1
Summary of trips recorded during data collection.

Time of day	Number of trip starts	Average trip duration	Range of trip duration
00:00–03:00	3	01:12:00	00:35:00
00:03–06:00	3	01:40:20	02:21:00
06:00–09:00	30	00:50:24	03:15:00
09:00–12:00	75	00:54:28	03:24:00
12:00–15:00	75	00:55:33	02:23:00
15:00–18:00	39	00:48:48	02:27:00
18:00–21:00	4	00:27:45	00:28:00
21:00–00:00	1	00:35:00	–
Total	230 trips	00:53:36	–

4.3. Locations at which transfers are made

The locations at which modal transfers were made were identified as the sections classified as WALK, WAIT, or a combination of both. To investigate the spatial distribution of these transfer locations, we mapped these locations in ArcGIS Pro. The intention is not to provide a

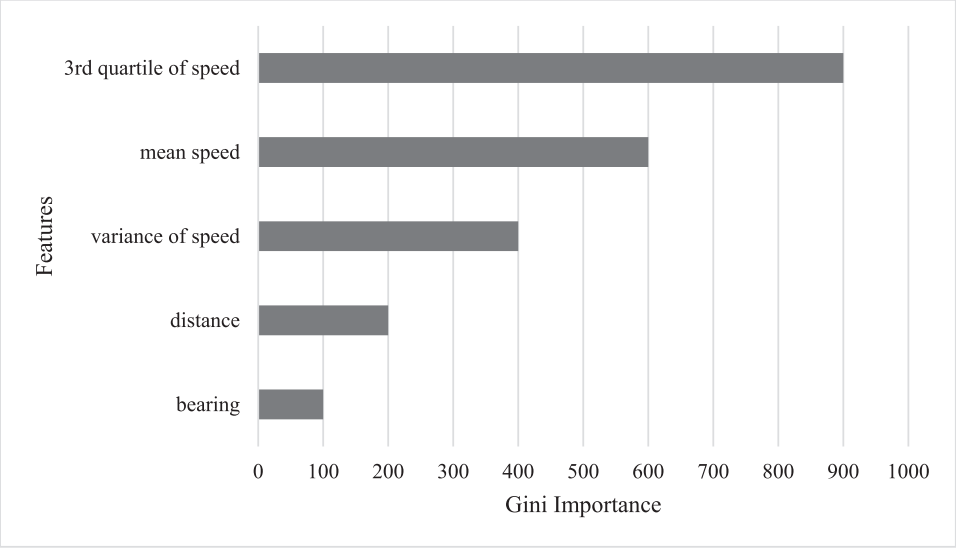

Fig. 4. Gini importance of extracted features in predicting the class of each trip segment.

Table 2
Confusion matrix for model testing.

Prediction	Reference		
	WALK	WAIT	RIDE
WALK	724	0	11
WAIT	4	483	0
RIDE	7	2	831

representative picture of transfer activity across the city (the sample is too small for this), but to investigate whether useful outputs can be generated with the approach. The visualisation was done at two levels of detail: at a macro level, to identify common transfer locations across the city, and at a micro level, to investigate passengers' transfer behaviour within specific precincts. Fig. 8 shows the citywide distribution of locations, using a heatmap plot that calculates the density of points on a raster layer to elevate areas with more transfer points in close proximity to each other. The figure also shows the network of taxi routes, mapped from a previous study (De Beer and Venter, 2024).

It is clear that transfers are not spread uniformly across the city but concentrated in a few areas:

Most transfers occur in the Pretoria Central Business District (CBD), which is consistent with the largely radial route pattern of minibus-taxi services. It is evident that many commuters travelling to and from areas outside the CBD are forced to transfer in the CBD.

The smaller community node of Garankuwa emerges as the second most common transfer location. Further investigation showed that this is at a local hospital, where a small formal taxi rank is provided to accommodate passengers visiting the hospital. It is interesting that this location has developed into a popular transfer location for the whole area. Spatially this location is at the edge of the Garankuwa township, at a natural connection point between local collector/distributor taxi services and line-haul routes connecting the area to more distant destinations like the CBD. The relative importance of this rank as a transfer location would not have been evident from any other data sources.

A number of other minor concentrations of transfers are spread across different areas in outlying residential areas in Garankuwa, Mamelodi, and Hammanskraal. Transfers are concentrated along main roads, at rail stations, and at junctions between taxis routes. Like in the previous case, many of these transfers likely occur between local routes, or between local and long-haul taxi services.

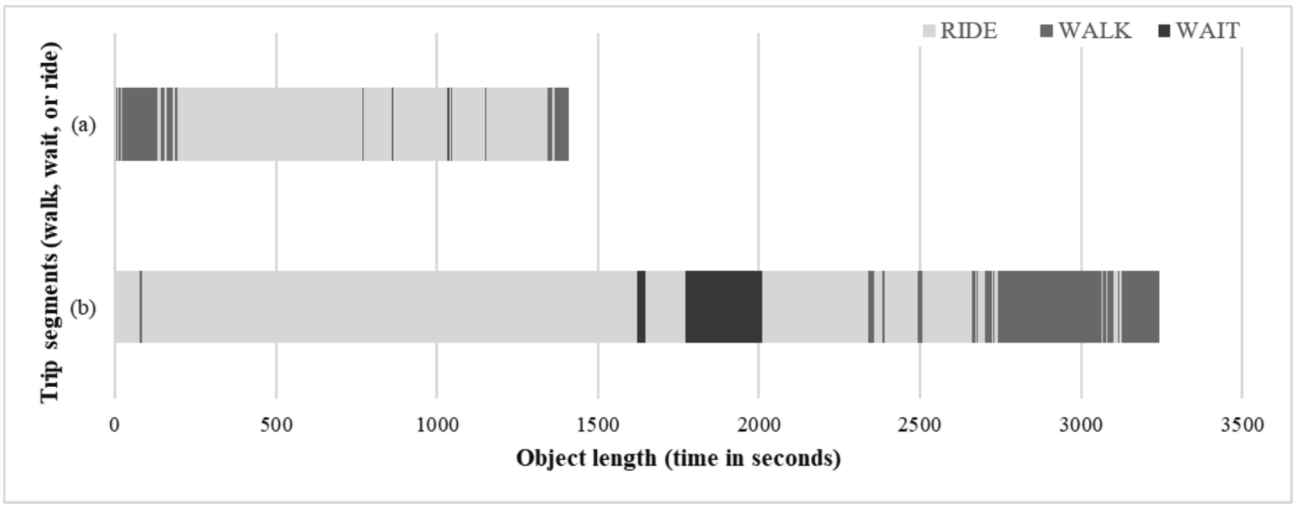

Fig. 5. Examples of trip segments for two trips as classified using the random forest classifier model: (a) trip with no transfer; (b) trip with transfer between 1700 and 2000 s.

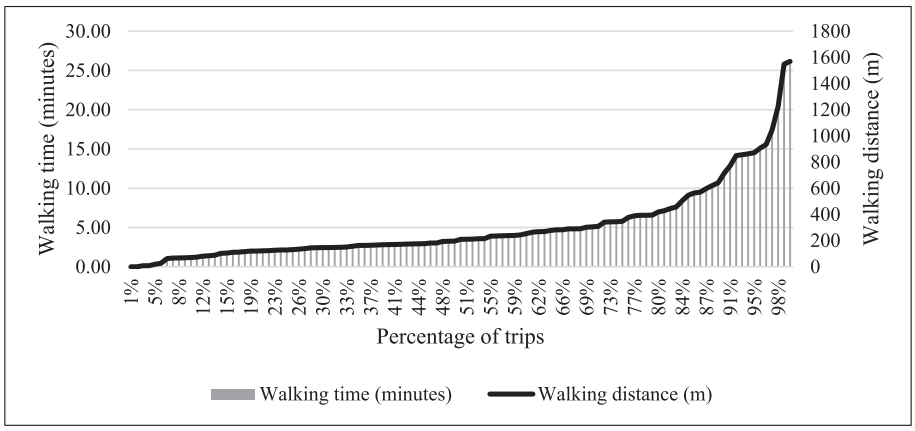

Fig. 6. Cumulative distribution of estimated transfer walking time (minutes).

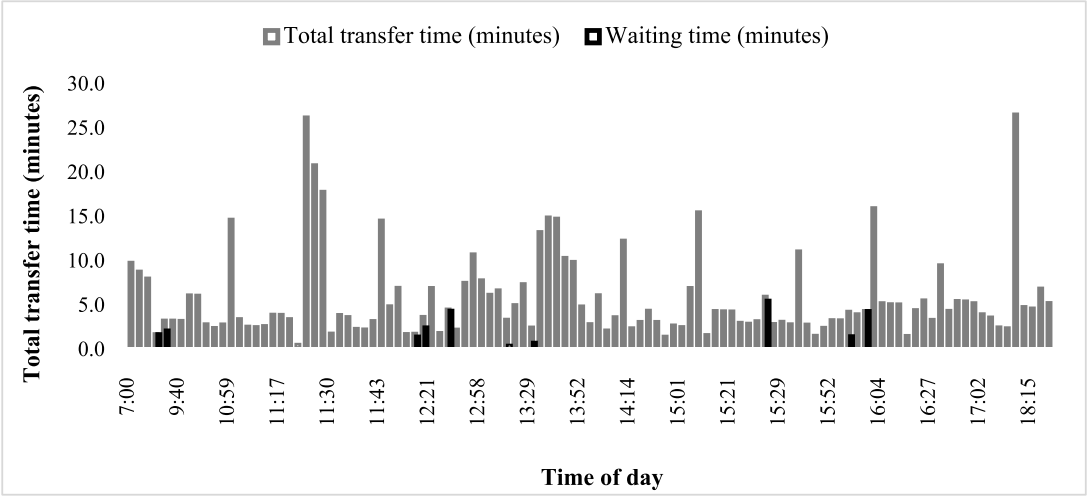

Fig. 7. Distribution of the waiting and total transfer time (minutes) by time of day.

Fig. 9 shows a microscale plot of all 18 identified walking segments within the Pretoria Central Business District, superimposed on a map of all formal and informal taxi facilities recorded by the municipality (City of Tshwane, 2015). Formal interchanges may include off-road undercover boarding areas and passenger amenities (Fig. 10 (a)), while informal facilities lack any infrastructure, and might be located in or next to the road reserve (Fig. 10 (b)). Informal facilities are established by operators and drivers without the intervention of authorities, usually in response to an undersupply of formal facilities in an area. From Fig. 9, a few points are salient. Firstly, passenger walks and waits occur across a wide stretch of the CBD. The flexibility of minibus-taxi operational patterns is such that passengers board and alight at many places near the route end points, and not just at designated facilities. Secondly, a considerable amount of walking subsequently takes place between taxis – some of the walk segments plotted are as long as 1.6 km. However, from visual observation it also appears that passengers engage in a variety of activities while walking, including buying food at roadside stalls and socialising. It is thus likely that passengers derive some utility from the transfer trip itself, which would complicate attempts to eliminate it.

Thirdly, several of the walk trips seem to start and end randomly along the same streets, at places where taxis occupy the parking lane and even park on sidewalks and medians (Fig. 10 (c)). This leads to what we term here "ribbon transferring", where entire street segments act as transfer locations. It is in most cases the result of inadequate transfer and storage facilities being provided for minibus taxis, and often associated with chaotic parking patterns, invasion of pedestrian spaces, and safety problems for pedestrians. Ribbon transferring therefore tends to take place under poor conditions for passengers and may lead to prolonged walk distances (compared to if transfers were contained within concentrated transfer facilities). We merely observe the phenomenon here and recognise that it needs further study.

5. Discussion

These observations suggest a number of implications for planning. Firstly, the evidence suggests that many transfers are concentrated in specific (fairly obvious) locations, such as in central areas where many radial routes converge. If the intention is to improve service quality for the passenger and multimodal integration of the public transport network, the evidence supports efforts by some authorities to invest in establishing and upgrading formal interchange facilities with adequate passenger amenities.

However, we also found that transferring in informal networks may encompass a diverse range of behaviours across a wide range of locations. These locations might include areas adjacent to (and between) formal ranks and stations, a myriad of minor transfer points within and between residential areas, and entire road sections where passengers engage in "ribbon" transferring between vehicles in the road reserve. Clearly, a better view is needed of this diversity, and the methods outlined in this study might help to identify these locations more clearly from actual passenger movement patterns. But the diffuse nature of this observed behaviour also raises the importance of general upgrading of

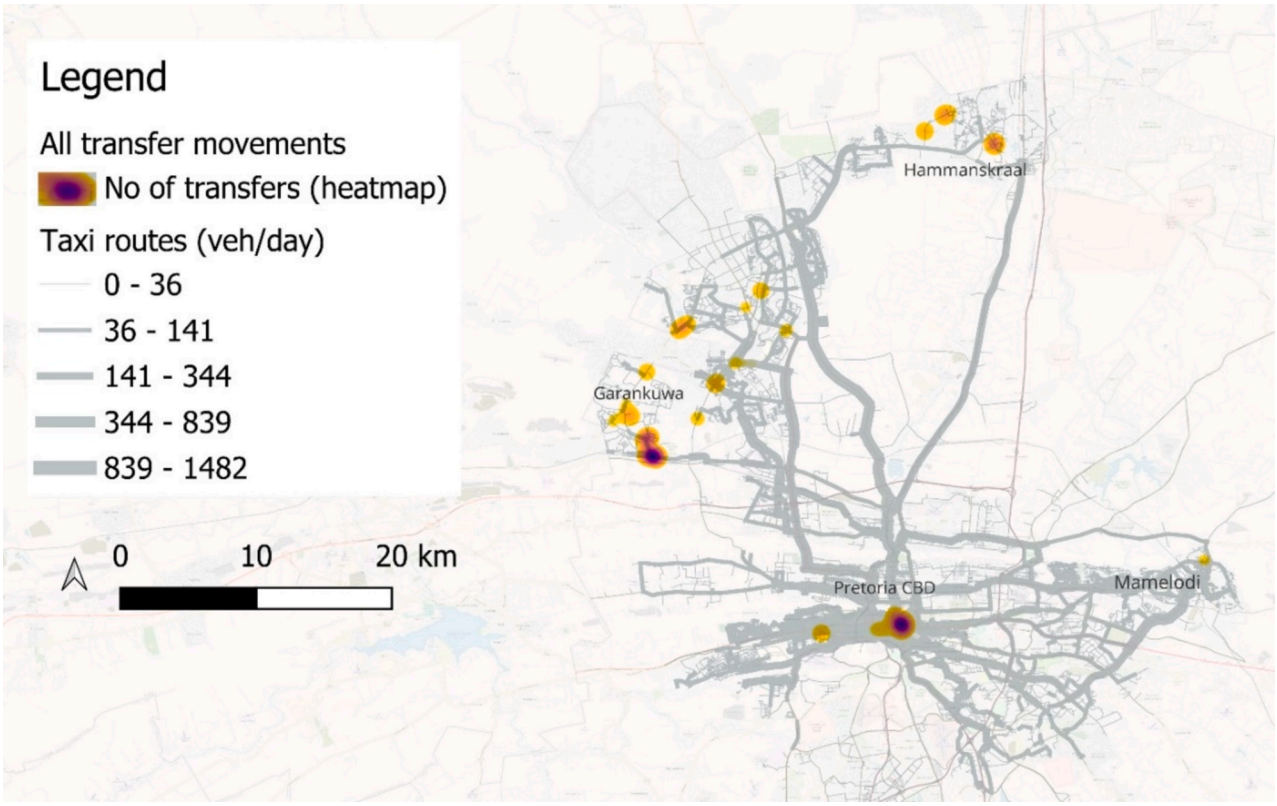

Fig. 8. Macroscale transfer walking and waiting segments, with taxi route network.

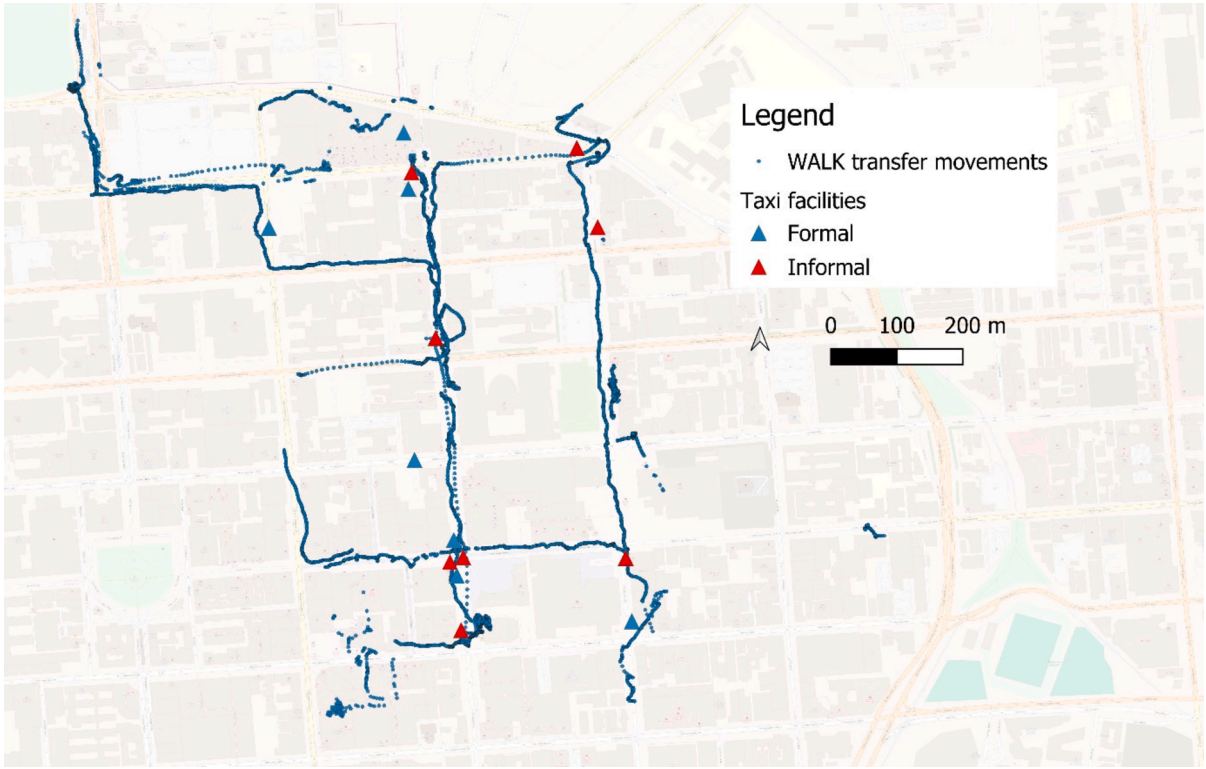

Fig. 9. Microscale plot of transfer walking segments and taxi facilities, Pretoria Central Business District. (Sources: Own data, and City of Tshwane, 2015).

the walkability and safety of the street environment in these areas for pedestrians, as a strategy for improving the transfer experience. To put it differently, targeted interventions to improve pedestrian service quality are critical to support the development of well-functioning and integrated multimodal networks, especially in informal networks whose connectivity is not optimised from a passenger perspective.

Fig. 10. Examples of (a) formal taxi transfer facility, (b) informal transfer locations without any formal infrastructure, and (b) linear or "ribbon" transfer outside a shopping mall. (Photos: O Mokoena; G Ankunda).

Secondly, the operating patterns of informal services are constantly evolving in response to changes in demand and industry conditions. An implication is that routes and transfer patterns will constantly change, leading to shifts in transfer locations. We documented the case above where a small community taxi rank evolved into a transfer node with regional importance, but without adequate capacity to handle the ensuing traffic volumes. Such shifts require authorities to adopt agile approaches to infrastructure provision, while striving to find the balance between pro-actively planning for and responding to evolving industry practice.

Lastly, transferring in informal networks seems to provide opportunities for a variety of social and commercial activities. Some of these become entrenched through the development of informal entrepreneurial businesses, for instance in small-scale food production or personal services. Authorities should take note of these activities, since they form part of the ecosystem of transferring that will not simply disappear with any attempts at formalisation or regularisation of network facilities. The intersection of transportation and bottom-up land use development represents a form of bottom-up Transit Oriented Development (TOD), and we encourage further research to leverage such initiatives as time and emissions-saving strategies. Studies of passenger transfer needs, behaviour, and perceptions such as Hernandez and Monzon (2016) can be useful for policy makers to determine the infrastructural needs for upgrading informal transfer locations to formal facilities or improving existing facilities.

6. Conclusions and recommendations

The study demonstrated that volunteered GPS data is a feasible data source that can help planners understand the transfer experience in multimodal networks in data-poor environments. The quantity of data is such that machine learning is a suitable approach to cleaning and processing it, leading to adequate models that can be useful in identifying transfer activities and studying their characteristics. The models represent a novel way of disaggregating transfer times into walking and waiting components.

Despite the positive outcomes of the proof-of-concept study, the research highlighted several ways in which the approach can be improved in future, potentially scaled up, studies. Further work is needed to refine the trip end identification algorithm, for instance by incorporating metrics other than stop duration. The random forest classifier used to distinguish between walking, waiting, and in-vehicle travel segments of a trip can be improved through collecting bigger datasets for training. The classification task is made more difficult by the diverse range of behaviours that are included in transferring, which includes information seeking, browsing, pacing, socialising, and shopping. This makes the distinction between walking and waiting behaviours during transfers less than clear-cut from a phenomenological perspective and raises important questions of how to introduce nuance into what is considered wait time. In addition, some waiting occurs inside vehicles, when passengers wait for a taxi to fill up sufficiently before the driver starts the trip. Our approach which includes both low-speed and stationary components in the definition of transferring behaviour seems to be reasonable, but the issue needs further exploration and clarification.

Behaviourally, we found that transferring typically involves more walking than waiting. Walking times are generally short, with a mean of just over 5 minutes, suggesting that most transfers are spatially efficient and that taxi operators, despite being self-organising, do not impose unreasonably long walking distances. This is consistent with the findings of Mittal et al. (2024), and might mean that they are aware of proximity of stopping points for the sake of passengers, or that they are driven by competitive behaviour because overly long walks introduce more opportunities for competing drivers to intercept the market. Nevertheless, a minority of transferring passengers are observed to experience longer transfer trips; a follow-up analysis might drill more deeply into causes for this variation.

The granularity of the GPS data allows the identification of transfer locations both at a macro and a micro scale. We find evidence of a large number and variety of transfer locations in different contexts. Transfer locations vary from major nodes in the network like the city centre where most line-haul taxi routes converge, to a wide range of minor

locations serving transfers between local taxi routes or local and line-haul services. We identify for the first time a type of linear or "ribbon" transferring along stretches of roads where many taxis use the road reserve for passenger operations; these may be a source of longer walking distances. Most of these transfer locations have no passenger facilities, making it difficult for authorities to know how and where to provide better infrastructure. We suggest that the methods outlined in this study might help to identify these locations more clearly from actual passenger movement patterns, before more in-depth assessments of the physical and contextual conditions at these locations are undertaken (Park and Chowdhury, 2022, Park and Chowdhury, 2018).

This work can be extended in a few directions. The collection of volunteered GPS data from a larger sample of passengers needs exploration, as this is needed to form a more representative picture of network-wide integration and transferring conditions. Our work and that of others (Yan et al., 2020, Howe, 2021) suggest that this might be challenging, but not impossible. The methods may also be extended to estimate network-wide metrics of multimodal integration and connectivity, along the lines of the work by Chowdhury et al. (2014). Data of this nature can be further utilised by disaggregating the characteristics of the different categories of transfers throughout the course of the day, particularly in the peak and off-peak periods.

Credit authorship contribution statement

Genevivie Ankunda: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Christo Venter:** Writing – review & editing, Visualization, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

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.

Acknowledgement

This research was funded by the Centre for Transport Development at the University of Pretoria. G. Ankunda also received partial support from the Mastercard Foundation Scholarship Programme at the University of Pretoria.

References

Aiswarya, K., Sriram, A., Raja, E. & Gandhimathi, G. An innovative scheme for smart school bus tracking system using machine learning and IoT techniques. AIP Conference Proceedings, 2023. AIP Publishing.
Alpaydin, E., 2020. Introduction to Machine Learning. MIT Press.
Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A. S., Al-Dabbagh, B. S. N., Fadhel, M. A., Manoufali, M., Zhang, J., Al-Timemy, A. H., Duan, Y., Abdullah, A., Farhan, L., Lu, Y., Gupta, A., Albu, F., Abbosh, A. & Gu, Y. 2023. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data,* 10.
Asakura, Y., Utsunomiya, Y., Hato, E. Professor, A. 2003. Verification of Stay and Move Identification Algorithm for Mobile Objects Using Observed Location Positioning Data.
Asakura, Y., Hato, E. Maruyama, T. 2014. Behavioural data collection using mobile phones. *In:* Rasouli, S. & Timmermans, H. (eds.) *Mobile Technologies for Activity-Travel Data Collection and Analysis.* United States of America: Infromation Science Reference (an imprint of IGI Global).
Auld, J., Mohammadian, A.K., 2014. Collecting activity-travel and planning process data using GPS-based prompted recall surveys: recent experience and future directions. Mobile Technologies for Activity-Travel Data Collection and Analysis. IGI Global.
Aziz, A., Nawaz, M., Nadeem, M., Afzal, L., 2018. Examining suitability of the integrated public transport system: a case study of Lahore. Transp. Res. A Policy Pract. 117, 13–25.
Basso, F., Feijoo, F., Pezoa, R., Varas, M., Vidal, B., 2023. The impact of electromobility in public transport: An estimation of energy consumption using disaggregated data in Santiago, Chile. Energy 129550.
Behrens, R., Mfinanga, D.A., McCormick, D., 2015. Paratransit in African Cities: Operations, Regulation and Reform. Routledge.
Behrens, R., Chalermpong, S., Oviedo, D., 2021. Informal paratransit in the Global South. The routledge handbook of public transport. Routledge.
Biau, G., Scornet, E., 2016. A random forest guided tour. TEST 25, 197–227.
Blazquez, C. A. & Miranda, P. A. 2015. A real time topological map matching methodology for gps/gis-based travel behavior studies. *Transportation Systems and Engineering: Concepts, Methodologies, Tools, and Applications.* IGI Global.
Bolbol, A., Cheng, T., Tsapakis, I., Haworth, J., 2012. Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. Comput. Environ. Urban Syst. 36, 526–537.
Breiman, L., 2001. Random forests. *Machine Learning* 45, 5–32.
Bricka, S. G., Simek, C. L. & Wood, N. 2014. Origin-destination data collection technology. *Mobile technologies for activity-travel data collection and analysis.* IGI Global.
Brown, G., 2017. A Review of Sampling Effects and Response Bias in Internet Participatory Mapping (PPGIS/PGIS/VGI). Trans. GIS 21, 39–56.
Ceder, A., 2021. Urban mobility and public transport: future perspectives and review. Int. J. Urban Sci. 25, 455–479.
Ceder, A., Le Net, Y., Coriat, C., 2009. Measuring public transport connectivity performance applied in Auckland, New Zealand. Transp. Res. Rec. 2111, 139–147.
Ceder, A. 2007. Public Transit Planning and Operation Theory, modelling and practice. Civil and Environmental Faculty. *Transportation Research Institute, Technion–Israel Institute of Technology, Haifa.*
Cervero, R., Golub, A., 2007. Informal transport: A global perspective. Transp. Policy 14, 445–457.
Ching, A., Zegras, C., Kennedy, S., Mamun, M., 2013. A user-flocksourced bus experiment in Dhaka: New data collection technique with smartphones. Transport. Res. Record: J. Transport. Res. Board.
Chowdhury, S., Ceder, A., Velty, B., 2014. Measuring public-transport network connectivity using Google Transit with comparison across cities. J. Public Transp. 17, 6.
Ciregan, D., Meier, U. & Schmidhuber, J. Multi-column deep neural networks for image classification. 2012 IEEE conference on computer vision and pattern recognition, 2012. IEEE, 3642-3649.
City of Tshwane, 2023. 2023–2024 Review of the 2022–2026 Integrated Development Plan. Pretoria, South Africa.
City of Tshwane 2013. Metropolitan Spatial Development Framework. Pretoria.
City of Tshwane 2015. Comprehensive Integrated Transport Plan 2015-2020. Pretoria.
Coetzee, J., Krogscheepers, C. Spotten, J. Mapping minibus-taxi operations at a metropolitan scale -methodologies for unprecedented data collection using a smartphone application and data management techniques. The 37th Southern African Transport Conference, 9 - 12 July 2018 2018 Pretoria, South Africa.
Costa, M.A., Marra, A.D., Corman, F., 2023. Public transport commuting analytics: a longitudinal study based on GPS tracking and unsupervised learning. Data Sci. Transport. 5, 15.
Dabiri, S., Heaslip, K., 2018. Inferring transportation modes from GPS trajectories using a convolutional neural network. Transp. Res. Part C Emerging Technol. 86, 360–371.
De Beer, L. & Venter, C. 2024. Using GPS data to determine minibus taxi driving behaviour and patterns. *First African Transport Research Conference.* Cape Town.
Techie Delight. 2023. *Run Length Encoding (RLE) Data Compression Algorithm* [Online]. Available: https://www.techiedelight.com/run-length-encoding-rle-data-compression-algorithm/ [Accessed 11 June 2023].
Department of Transport 2016. National Transport Master Plan (NATMAP) 2050 Synopsis Report. *In:* Department of Transport (ed.). Pretoria, South Africa.
Du Preez, D., Zuidgeest, M., Behrens, R., 2019. A quantitative clustering analysis of paratransit route typology and operating attributes in Cape Town. J. Transp. Geogr. 80, 102493.
Ehrlich, J., Hard, E., Komanduri, A. & Anderson, R. S. 2020. A Century of Travel Surveys Informing Transportation Investments. *Centennial Papers.*
Falchetta, G., Noussan, M., Hammad, A., 2021. Comparing paratransit in seven major African cities: An accessibility and network analysis. J. Transp. Geogr. 94, 103131.
Fan, J., Fu, C., Stewart, K., Zhang, L., 2019. Using big GPS trajectory data analytics for vehicle miles traveled estimation. Transp. Res. Part C Emerging Technol. 103, 298–307.
Fang, K. & Zimmerman, S. 2015. Public Transport Service Optimization and System Integration. China Transport Topics, No. 14;. Washington, DC: World Bank.
Fang, S.-H., Fei, Y.-X., Xu, Z., Tsao, Y., 2017. Learning transportation modes from smartphone sensors based on deep neural network. IEEE Sens. J. 17, 6111–6118.
Feng, T., Timmermans, H., 2014. Multi-week travel surveys using GPS devices: experiences in The Netherlands. Mobile technologies for activity-travel data collection and analysis. IGI Global.
Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2014. Do we need hundreds of classifiers to solve real world classification problems? J. Machine Learn. Res. 15, 3133–3181.
Ferrer, S., Ruiz, T., 2014. Using smartphones to capture personal travel behavior. In: Timmermans, H., Soora, R. (Eds.), Mobile Technologies for Activity-Travel Data Collection and Analysis. IGI Global.
Ferro, P.S., 2015. Paratransit: A Key Element in a Dual System. Agence Française de Développement, Paris.
Ferster, C. J., Nelson, T., Robertson, C. & Feick, R. 2018. 1.04 - Current Themes in Volunteered Geographic Information. *In:* Huang, B. (ed.) *Comprehensive Geographic Information Systems.* Oxford: Elsevier.
Garcia-Martinez, A., Cascajo, R., Jara-Diaz, S.R., Chowdhury, S., Monzon, A., 2018. Transfer penalties in multimodal public transport networks. Transp. Res. A Policy Pract. 114, 52–66.

Goenaga, B., Underwood, B.S., Castorena, C., Cantillo, V., Arellana, J., 2023. Using continuous traffic counts extracted from smartphone data to evaluate traffic reductions during COVID-19 pandemic in North Carolina. Latin Am. Transport Stud. 1, 100005.
Gong, L., Sato, H., Yamamoto, T., Miwa, T., Morikawa, T., 2015. Identification of activity stop locations in GPS trajectories by density-based clustering method combined with support vector machines. J. Modern Transport. 23, 202–213.
Goodchild, M.F., 2007. Citizens as sensors: the world of volunteered geography. GeoJournal 69, 211–221.
Gotz, G., Wray, C., Venter, C., Badenhorst, W., Trangoš, G., Culwick, C., 2015. Mobility in the Gauteng City-Region. Gauteng City-Region Observatory (GCRO), Johannesburg.
Guo, Z., Wilson, N.H., 2011. Assessing the cost of transfer inconvenience in public transport systems: A case study of the London Underground. Transp. Res. A Policy Pract. 45, 91–104.
Han, H., Guo, X. & Yu, H. Variable selection using mean decrease accuracy and mean decrease gini based on random forest. 2016 7th ieee international conference on software engineering and service science (icsess), 2016. IEEE, 219-224.
Han, S., Williamson, B.D., Fong, Y., 2021. Improving random forest predictions in small datasets from two-phase sampling designs. BMC Med. Inf. Decis. Making 21, 1–9.
Hayduk, B. W. 1997. Multimodal Transportation Planning Data: Compendium of Data Collection Practices and Sources.
Hayes, G. & Venter, C. An innovative method to collect route choice preference data using a smartphone application. 2022. Southern African Transport Conference.
Hernandez, S., Monzon, A., 2016. Key factors for defining an efficient urban transport interchange: Users' perceptions. Cities 50, 158–167.
Hitge, G., Vanderschuren, M., 2015. Comparison of travel time between private car and public transport in Cape Town. J. South Afr. Inst. Civil Eng. 57, 35–43.
Howe, J., 2008. Crowdsourcing: How the Power of the Crowd is Driving the Future of Business. Random House.
Howe, L.B., 2021. Thinking through people: The potential of volunteered geographic information for mobility and urban studies. Urban Stud. 0042098020982251.
Icasa, 2020. The State of the ICT Sector Report in South Africa. Pretoria, South Africa.
Institute for Transportation & Development Policy (ITDP) 2016. The BRT Standard. *Access and Integration.*
Janiesch, C., Zschech, P., Heinrich, K., 2021. Machine learning and deep learning. Electron. Mark. 31, 685–695.
Jianchuan, X., Zhicai, J., Guangnian, X. & Xuemei, F. 2014. Smartphone-Based Travel Survey: A Pilot Study in China. *In:* Rasouli, S. & Timmermans, H. (eds.) *Mobile Technologies for Activity-Travel Data Collection and Analysis.* United States of America: Information Science Referebce (an imprint of IGI Global).
Joseph, L., Neven, A., Martens, K., Kweka, O., Wets, G., Janssens, D., 2020. Measuring individuals' travel behaviour by use of a GPS-based smartphone application in Dar es Salaam, Tanzania. *J. Transport Geogr.* 88, 102477.
Kash, G., Hidalgo, D., 2014. The promise and challenges of integrating public transportation in Bogotá, Colombia. Public Transp. 6, 107–135.
Kerzhner, T., 2023. How are informal transport networks formed? Bridging planning and political economy of labour. Cities 137.
Klopp, J.M., Cavoli, C.M., 2017. The paratransit puzzle: Mapping and master planning for transportation in Maputo and Nairobi. *Urban mobilities in the global south.* Routledge.
Klopp, J., Williams, S., Waiganjo, P., Orwa, D. & White, A. 2015. Leveraging cellphones for wayfinding and journey planning in semi-formal bus systems: Lessons from digital matatus in Nairobi. *In:* Geertman, S., Ferreira, J., J., Goodspeed, R. & Stillwell, J. (eds.) *Planning support systems and smart cities.* Cham: Springer.
Kohla, B., Gerike, R., Hössinger, R., Meschik, M., Sammer, G., Unbehaun, W., 2014. A new algorithm for mode detection in travel surveys: mobile technologies for activity-travel data collection and analysis. Mobile technologies for activity-travel data collection and analysis. IGI Global.
Krygsman, S. C. Nel, J. 2009. The use of global positioning devices in travel surveys-a developing country application. *SATC 2009.*
Kumar, A. M., Zimmerman, S. Arroyo Arroyo, F. Myths and Realities of Informal Public Transport in Developing Countries. 2021.
Lison, P., 2015. An introduction to machine learning. Language Technol. Group (LTG) 1, 1–35.
Manana, K. Progress with the Implementation of IPTNS in South African Cities. The 39th Annual Southern African Transport Conference and Exhibition, 5 - 7 July 2021 2021 Virtual Event.
Martinez-Taboada, F., Redondo, J.I., 2020. Variable importance plot (mean decrease accuracy and mean decrease Gini). PLOS ONE.
McCormick, D., Schalekamp, H., Mfinanga, D., 2015. The nature of paratransit operations. Paratransit in African Cities. Routledge.
McKay, T., Simpson, Z., Patel, N., 2017. Spatial politics and infrastructure development: analysis of historical transportation data in Gauteng-South Africa (1975–2003). Miscellanea Geographica 21, 35–43.
Mittal, K.M., Timme, M., Schröder, M., 2024. Efficient self-organization of informal public transport networks. Nat. Commun. 15, 4910.
Mokoma, L., Venter, C., 2023. Pathways to integrating paratransit and formal public transport: case studies from Tshwane, South Africa. *Res. Transport. Econ.* 102, 101356.
Molloy, J., Castro, A., Götschi, T., Schoeman, B., Tchervenkov, C., Tomic, U., Hintermann, B., Axhausen, K.W., 2023. The MOBIS dataset: a large GPS dataset of mobility behaviour in Switzerland. Transportation 50, 1983–2007.
Moodley, S., Venter, C., 2022. Measuring the service quality at multimodal public transport interchanges: a needs-driven approach. Transp. Res. Rec. 2676, 194–206.
Nasteski, V., 2017. An overview of the supervised machine learning methods. Horizons. b 4, 51–62.
Ndibatya, I., Booysen, M., 2020. Minibus taxis in Kampala's paratransit system: Operations, economics and efficiency. J. Transp. Geogr. 88, 102853.
Ndibatya, I., Booysen, M., 2021. Characterizing the movement patterns of minibus taxis in Kampala's paratransit system. J. Transp. Geogr. 92, 103001.
Ndibatya, I., Coetzee, J., Booysen, M., 2017. Mapping the informal Public Transport network in Kampala with smartphones. Civil Eng. -Siviele Ingenieurswese South African Institut. Civil Eng. (SAICE) 1, 35–40.
NEA, OGM, TSU, 2003. Integration and regulatory structures in public transport. European Commission DG TREN, Brussels.
Ouyang, W. & Wang, X. Joint deep learning for pedestrian detection. Proceedings of the IEEE international conference on computer vision, 2013. 2056-2063.
Palencia Arreola, D.H., 2019. Arguments for and field experiments in democratizing digital data collection: the case of Flocktracker. Massachusetts Institute of Technology.
Park, J., Chowdhury, S., 2018. Investigating the barriers in a typical journey by public transport users with disabilities. J. Transp. Health 10, 361–368.
Park, J., Chowdhury, S., 2022. Towards an enabled journey: barriers encountered by public transport riders with disabilities for the whole journey chain. Transp. Rev. 42, 181–203.
Parmar, A., Katariya, R. & Patel, V. A review on random forest: An ensemble classifier. International conference on intelligent data communication technologies and internet of things (ICICI) 2018, 2019. Springer, 758-763.
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., 2011. Scikit-learn: Machine learning in Python. J. Machine Learning Res. 12, 2825–2830.
Plano, C., Behrens, R., 2022. Integrating para-and scheduled transit: Minibus paratransit operators' perspective on reform in Cape Town. Res. Transp. Bus. Manag. 42, 100664.
Plano, C., Behrens, R., Zuidgeest, M., 2020. Towards evening paratransit services to complement scheduled public transport in Cape Town: A driver attitudinal survey of alternative policy interventions. Transp. Res. A Policy Pract. 132, 273–289.
Reinau, H. K., Harder, H. & Overgard, H. C. 2014. Horses for Courses: Designing a GPS Tracking Data Collection. *In:* Rasouli, S. & Timmermans, H. (eds.) *Mobile Technologies for Activity-Travel Data Collection and Analysis.* United States of America: Infromation Science Reference (an imprint of IGI Global).
Rodrigue, J.-P. 2024. Transportation and Spatial Structure. *The Geography of Transport Systems.* Sixth Edition ed.: Routledge.
Saddier, S. & Johnson, A. 2018. Understanding the operational characteristics of paratransit services in Accra, Ghana: A case study.
Schakenbos, R., La Paix, L., Nijenstein, S., Geurs, K.T., 2016. Valuation of a transfer in a multimodal public transport trip. Transp. Policy 46, 72–81.
Statistics South Africa 2022. National Household Travel Survey, 2020. *In:* Statistics South Africa (ed.). Pretoria, South Africa.
Stopher, P.R., 2009a. Collecting and processing data from mobile technologies. In: Bonnel, P., Lee-Gosselin, M., Zmud, J., Madre, J.-L. (Eds.), Transport Survey Methods: Keeping up with a Changing World. Emerald Group Publishing Limited, UK.
Stopher, P.R., 2009b. The Travel Survey Toolkit: Where to From Here? Transport Survey Methods: Keeping up With a Changing World. Emerald Group Publishing Limited.
Thin, L.N., Ting, L.Y., Husna, N.A., Husin, M.H., 2016. GPS systems literature: inaccuracy factors and effective solutions. Int. J. Computer Networks Commun. (IJCNC) 8, 123–131.
Tun, T. H., Welle, B., Hidalgo, D., Albuquerque, C., Castellanos, S., Sclar, R. Escalante, D. 2020. Informal and semiformal services in Latin America: an overview of public transportation reforms.
Varghese, V., Chikaraishi, M., Urata, J., 2020. Deep learning in transport studies: A meta-analysis on the prediction accuracy. J. Big Data Anal. Transport. 2, 199–220.
Venter, C., Minora, N., Shukrani, K. & du Toit, J. 2014. A Role for GPS Data in Qualitative Research: Exploring Links between Walking Behaviour, the Built Environment, and Crime Perception in South Africa. *In:* Rasouli, S. & Timmermans, H. (eds.) *Mobile Technologies for Activity-Travel Data Collection and Analysis.* United States of America: Information Science Reference (an imprint of IGI Global).
Yan, Y., Feng, C.-C., Huang, W., Fan, H., Wang, Y.-C., Zipf, A., 2020. Volunteered geographic information research in the first decade: a narrative review of selected journal articles in GIScience. Int. J. Geogr. Inf. Sci. 34, 1765–1791.
Yang, Y., Jia, B., Yan, X.-Y., Li, J., Yang, Z., Gao, Z., 2022. Identifying intercity freight trip ends of heavy trucks from GPS data. Transport. Res. Part e: Logist. Transport. Rev. 157, 102590.
Yazdizadeh, A., Patterson, Z., Farooq, B., 2019. An automated approach from GPS traces to complete trip information. Int. J. Transp. Sci. Technol. 8, 82–100.
Yun, H.Y., Zegras, C., Palencia Arreola, D.H., 2019. "Digitalizing walkability": Comparing smartphone-based and web-based approaches to measuring neighborhood walkability in Singapore. J. Urban Technol. 26, 3–43.
Zheng, Y., Chen, Y., Li, Q., Xie, X., Ma, W.-Y., 2010. Understanding transportation modes based on GPS data for web applications. ACM Trans. Web 4, 1–36.