

Developing a representative driving cycle for paratransit that reflects measured data transients: Case study in Stellenbosch, South Africa

Christopher Hull ^{*}, Katherine A. Collett, Malcolm D. McCulloch

Energy and Power Group at Engineering Science Department, University of Oxford, OX2 0ES, United Kingdom

ARTICLE INFO

Keywords:

Transportation planning
Urban development
GPS data
Clustering
Time series
Vehicle design

ABSTRACT

Paratransit plays a critical role in meeting transportation needs in many cities in sub-Saharan Africa (SSA). However, it faces deep issues related to pollution, congestion, and safety. Understanding the driving patterns of paratransit in SSA can provide valuable insights into the transportation needs in the region, which is particularly relevant nowadays given the increasing focus on sustainable transportation solutions in Africa.

Representative driving cycles, which provide a realistic simulation of the driving conditions a vehicle is likely to encounter, are key to framing policies for effective transportation management, vehicle design, and urban and regional planning. However, cycle development has been limited in SSA due to a lack of data and standardized testing procedures. This study develops a representative driving cycle using GPS data gathered on paratransit vehicles traveling around Stellenbosch, South Africa, providing a benchmark for evaluation and a platform for further research and testing in SSA's dominant transport industry.

A novel time series shape-based clustering methodology is employed that combines dynamic time warping and mixed integer programming to cluster micro-trips of varying length based on their time series shapes. Representative micro-trips from each cluster are stitched together with a maximum likelihood approach to curate the final cycle. By including transients from the measured data in cycle development, this novel approach to cycle development is particularly suited for capturing the notoriously unconventional and aggressive driving style of paratransit.

The constructed cycle and several international cycles are assessed against the measured database on the basis of eight characteristic kinematic parameters. The constructed cycle emerges as the most fitting choice to represent paratransit operating conditions, with an average deviance of 3.65% across the parameters, compared to deviations of 23%–34% for the international cycles.

1. Introduction

Transportation systems in sub-Saharan Africa (SSA) face significant challenges related to pollution, congestion, and safety. Paratransit, which describes the informal, demand-driven, privately owned transportation systems that dominate road transport throughout SSA (McCormick et al., 2016; Evans et al., 2018), suffers from these issues even while playing a critical role in meeting the transportation needs of the population. Its high affordability, coverage, and flexibility, making it accessible to citizens living in a wide range of socioeconomic classes and locations, and it provides over 70% and up to 98% of public transport ridership in many

^{*} Corresponding author.

E-mail address: christopher.hull@eng.ox.ac.uk (C. Hull).

<https://doi.org/10.1016/j.tra.2024.103987>

Received 11 February 2023; Received in revised form 27 November 2023; Accepted 22 January 2024

Available online 30 January 2024

0965-8564/© 2024 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).



Fig. 1. The minibus taxi, most common paratransit vehicle in sub-Saharan Africa (Toyota, 2023).

major cities in the region (Behrens et al., 2015; Evans et al., 2018), including Dar es Salaam, Johannesburg, Lagos, and Kampala. The most common paratransit vehicle – accounting for roughly 83% of the paratransit sector's trips and servicing approximately 72% of daily commuters – is the minibus taxi, depicted in Fig. 1, which carries from 9 to 16 passengers, depending on the vehicle model (McCormick et al., 2016; Evans et al., 2018), and weighs 3000–4000 kg (Toyota, 0000).

However, the design and operation of paratransit systems can be hindered by a lack of accurate and representative driving cycles that reflect the unique driving conditions of the region (Hull et al., 2023; Giliomee et al., 2023; Collett and Hirmer, 2021). Driving cycles are vehicle speed-time series that simulate real-world driving conditions. They have a wide range of uses, from determining fuel consumption and emissions profile, to designing traffic control systems and determining the performance of vehicle systems (Arun et al., 2017; Tong et al., 2011; Badusha and Ghosh, 1999; Ho et al., 2014; Mayakuntla and Verma, 2018; Kamble et al., 2009). In practice, this can imply that driving cycles will dictate the procedure a vehicle is run through (often on a machine such as a chassis dynamometer) to measure performance under typical driving conditions. Cycles can also be used by engineers for vehicle technology research & development, or by policymakers and urban planners for transportation planning. Consequently, a detailed understanding of driving cycles can be useful for guiding long term vehicle design (Bishop et al., 2012; Wangsness et al., 2020) such as engine power, fuel type (e.g. diesel or electric), and vehicle size, constructing policies that efficiently manage congestion and transportation emissions (Aksen et al., 2020; Dixit and Sivakumar, 2020; Gupta and Garg, 2020; Jaikumar et al., 2017; Kim and Kim, 2020), and generally understanding transportation system operations.

It is particularly important that driving cycles accurately represent the driving patterns of the target vehicle type in the target region, so that results from performance tests capture realistic vehicle performance. However, well-known standard international driving cycles such as the New European Driving Cycle (NEDC) or Worldwide Harmonized Light Vehicle Procedure (WLTP) have been found in various studies and reviews on driving cycle studies to insufficiently represent local driving patterns in many developed economies (Tong and Ng, 2023; Chandrashekar et al., 2021; Ganesh Sundarkumar et al., 2021; Tong et al., 2011; Ho et al., 2014).

Over the past four to five decades, representative driving cycles have been constructed for cities around the world (e.g. Galgamuwa et al., 2015; Ho et al., 2014; Tong et al., 2011; Arun et al., 2017; Rodríguez et al., 2016). To date, it is clear from reviews of the driving cycle literature that the regions of focus for developing these cycles have been the USA, Europe, South America, and Asia (Galgamuwa et al., 2015; Quirama et al., 2020; Tong and Ng, 2023). While there have been some isolated forays into analyzing the energy consumption (Abraham et al., 2023, 2021; Booysen et al., 2022; Hull et al., 2023; Giliomee et al., 2023) and mobility (Ndibatya and Booysen, 2021, 2020b,a) of paratransit in SSA, the driving patterns for paratransit systems in SSA are not well studied. Limitations in SSA stem primarily from a lack of data on actual driving patterns and a lack of standardized testing procedures. By contrast, the driving cycles of electric passenger vehicles and larger buses in the developed world have been analyzed heavily (Brady and O'Mahony, 2016; Berzi et al., 2016; Kivekäs et al., 2018b; Smith et al., 2011; Wang et al., 2008).

The objective of this study are to develop a suitable methodology for constructing driving cycles specific to paratransit vehicles, and to create a representative paratransit driving cycle. To do so, it adopts a novel time series shape-based clustering methodology to develop a cycle from 1 Hz GPS tracking data collected on paratransit vehicles traveling around Stellenbosch, South Africa. This research contributes to the understanding of how paratransit systems operate and how they can be improved in SSA. The study's outcomes are expected to be useful for researchers, policymakers, and vehicle manufacturers to better understand and address the transportation challenges in the region.

Given the unique mobility patterns and unconventional driving styles of African paratransit (Ndibatya and Booysen, 2021; Zeeman and Booysen, 2014; Hull et al., 2023), the application of standard international, typically derived from high-income countries, prove unsuitable in these contexts. Although the cycle developed here is limited to Stellenbosch, the study presents a methodology that can be flexibly adapted for different geographies and driving conditions. The methodology is based on “micro-trip” clustering. Micro-trips are trip segments where the vehicle goes from stopped to stopped. Dynamic time warping is used to measure the similarity between micro-trips of different lengths based on their time series shape, and mixed integer programming is used to find a globally optimal clustering solution. This approach allows for the clustering of micro-trips of varying length, which has not been possible in previous proposed methodologies that cluster micro-trips on their time series shape (Ganesh Sundarkumar et al., 2021). Furthermore, this study introduces the parameter t_{\min} which restricts micro-trips to a minimum duration, thus avoiding the inclusion of arbitrarily small or ‘spam’ micro-trips in the clustering process. Micro-trips nearest to cluster centers are used as the basis for a representative cycle, and a maximum likelihood method is used to ensure these candidate micro-trips are sequenced in an order that is representative of what is expected to be seen on the road. The usage of unaltered micro-trips from the measured

data in the final cycle ensures that the cycle captures real-world transients from the data, which are inherently difficult to capture in statistical constructions of driving cycles. In this context, “data transients” refer to temporary and often abrupt variations or irregularities in the driving data collected from paratransit vehicles, such as sudden accelerations, decelerations, stops, or other unique driving patterns characteristic of the aggressive, stop-start nature of paratransit driving. It is essential to include these data transients in the construction of representative driving cycles for paratransit vehicles to provide a comprehensive and accurate reflection of the actual driving conditions and behaviors experienced by paratransit drivers.

The novel contributions of this study are threefold:

- Develops a representative driving cycle for paratransit in sub-Saharan Africa.
- Presents a methodology to cluster micro-trips of varying lengths based on their time series shape. The clustering approach finds the globally optimal clustering solution.
- Introduces the parameter t_{\min} , which represents the minimum duration for a trip segment to be considered a micro-trip.

The rest of the paper is organized as follows. Section 2 provides an overview of driving cycle classification and the state of the art in driving cycle construction. Section 3 steps through the proposed methodology for this study. Section 4 presents the developed cycles and assesses how well they represent the real-world measured data in comparison to several international cycles. Section 5 concludes the study and provides some research perspectives.

2. Literature review

This section provides an overview of how driving cycles are typically classified, and a brief state of the art on methods for developing representative driving cycles. It elucidates how the methodology presented in this study is an improvement upon previous driving cycle development methodologies.

2.1. Driving cycle classification

It is common in vehicle design and legislation to default to the utilization of standard international driving cycles, known as ‘legislative’ driving cycles. However in practice, differences in infrastructure condition, traffic conditions, city and geographical features, economic and cultural context, and vehicle type mean that real-world driving patterns and therefore driving cycles vary greatly between cities and regions (Kamble et al., 2009; Zhao et al., 2018). Consequently, standard international cycles neither accurately represent driving conditions encountered in the real world, nor capture context specific driving patterns and transients (Degraeuwe and Weiss, 2017; Tong and Ng, 2023). To truly understand the benefits of proposed legislative, design, or engineering changes, there is a need for real-world representative driving cycles that are based on driving data from vehicles in any given region of interest (Vámosi et al., 2022). This is especially salient for paratransit vehicles, which are known for their uniquely aggressive driving style, characterized by many quick stop/start maneuvers (Zeeman and Booyesen, 2014; Ndibatya and Booyesen, 2021; Giliomee et al., 2023).

In addition to the legislative versus non-legislative classifications, driving cycles are classified as either ‘transient’ or ‘modal’ depending on how they are developed. Transient cycles are developed from on-road driving data, whereas modal cycles are developed from a sequence of constant speed and acceleration states or ‘modes’. Examples of transient cycles include the Worldwide Harmonized Light Vehicles Test Cycle (WLTC), the Hong Kong Driving Cycle, and the Standard Federal Test Procedure (SFTP-US06). Examples of modal cycles include the New European Driving Cycle (NEDC) and Japan’s J10-J15 Driving Cycle (Ho et al., 2014).

The modal method is recognized in the literature as not accurately capturing real world driving behavior, since the final aggregated speed-time series is synthetically smoothed, and the effects of transient acceleration/deceleration events are lost (Esteves-Booth et al., 2001; Tsai et al., 2005; Nylund et al., 2007). Moreover, it has been estimated that modal driving cycles underestimate emissions by 30%–50% (Michel, 2004; Bishop et al., 2012). In general, transient driving cycles, which are based on real-world data, are superior to modal driving cycles for real-world applications (Ganesh Sundarkumar et al., 2021). For example, the growing importance of EVs has increased the emphasis on the development of representative real-world driving cycles that are useful for predicting battery health (Pfriem and Gauterin, 2016), which modal driving cycles would fail to do. Since this study uses real-world on road driving data for cycle construction, the representative driving cycles developed here are transient.

2.2. Cycle construction methodologies

In general, there are five main categories of methodologies for constructing transient driving cycles in the literature (Chen et al., 2022). Many of them involve ‘micro-trips’, which are traditionally defined in the literature as segments of a trip where the vehicle goes from stopped to stopped. Some studies have defined micro-trips as trip segments where engine RPM goes from zero to zero (Ganesh Sundarkumar et al., 2021), which can be useful when evaluating vehicle’s emissions performance directly from the driving cycle. However, engine RPM data was not available for this study, so vehicle speed is used.

The five categories are random selection-based methods (RS), segment-based (SB) methods, monte carlo markov chain (MCMC) based methods, clustering analysis (CA) methods, and optimization based (OB) methods. There are other studies that fall outside of these categories, but they typically still involve the usage of micro-trips and some statistical construction of driving cycles (Kaymaz et al., 2019).

1. Random selection-based methods (RS): These methods are simple to implement and flexible enough to be used for a variety of vehicles and locations. They involve splitting the measured database into a series of micro-trips, and randomly selecting micro-trips to stitch together that meet the assessment parameters until the target cycle duration is reached. Some papers that use it include the following: [Arun et al. \(2017\)](#) used this method to develop driving cycles for passenger cars and motorbikes in Chennai, India. [Kamble et al. \(2009\)](#) construct a representative driving cycle using data from five major roads in Pune, India. [Tong \(2019\)](#) develop driving cycles for both diesel and super-capacitor electric bus vehicles in Hong Kong. [Ho et al. \(2014\)](#) build a driving cycle for Singapore and demonstrate that it is more representative than the New European Driving Cycle (NEDC) for estimating emissions. [Saleh et al. \(2010\)](#) compare driving cycles developed in Edinburgh and Delhi build using the RS. [Tong et al. \(2011\)](#) develop the first set of driving cycles for Hanio using RS to match overall summary statistics. RS is well studied and widely used. They have a number of advantages, including its simplicity and ease of implementation. They are also a flexible method that can be used to develop driving cycles for a variety of vehicles and locations. However, a disadvantage is that it can be difficult to use this method generate driving cycles that are representative of all driving conditions. Extreme driving events, such as sudden accelerations, hard braking, or emergency maneuvers, may not be adequately represented in cycles generated by RS due to the random nature of the selection process. Additionally, the RS can be computationally expensive, especially for large datasets.
2. Segment-based methods (SB): Segment-based cycle construction resembles micro-trip-based cycle construction except that the selection of a trip “segment” takes into account factors such as roadway type or Level of Service rather than considering adjacent stops. Since trip segments are partitioned based on traffic conditions and the road’s physical attributes, a trip segment can commence and conclude at varying speeds. Consequently, a key step when integrating these trip segments into a driving cycle is to align the speed and acceleration between two successive connecting points within micro-trips. This method is more appropriate for constructing driving cycles for expressways where there may not be many stops along the route. Using SB, [Dai et al. \(2008\)](#) construct a driving cycle for arterial driving with data from urban areas around California, and find that their segment based method creates more accurate cycles for speed activity at the high (>47.5mph) and low (<12.5mph) ends in their data. [Kivekäs et al. \(2018a\)](#) apply the method to approach to develop a method that they use to analyze the impact of driving cycle variations on electric buses in Espoo, Finland. [Shen et al. \(2018\)](#) build a driving cycle using a bus-station-based and whole-trip-based segmenting method for hybrid electric bus in Shanghai, to avoid velocity fluctuations between segments, and find the error rate of the most characteristic parameters of the typical hybrid bus cycle to be within 5%. The main advantage of SB lies in the ability to derive a cycle that is tailored to specific driving conditions. However, SB are not suitable for vehicles in complex driving scenarios, or ones that operate in unpredictable driving environments.
3. Monte Carlo-Markov Chain (MCMC) methods: The basic premise of MCMC is to predict the future vehicle state based on current state. To do so, it uses maximum likelihood estimation to cluster velocity-time data into fragments based on parameters such as velocity and acceleration. These fragments are in turn used to create modal bins containing fragments with similar characteristics. The transition probabilities between modes are estimated and used to identify a transition path between modes or micro-segments. [Bishop et al. \(2012\)](#) offer a data-driven Markov chain method to construct a driving cycle without decomposing the raw velocity-time sequence of the vehicles. [Gong et al. \(2018\)](#) use high frequency data to develop a Beijing driving cycle for battery electric vehicles, which is demonstrated to be more consistent with real-world Beijing driving than the standard international driving cycles. [Li et al. \(2017\)](#) propose a multi-scale single step prediction method that improves the prediction accuracy of the MCMC method by 7.18% on average. [Liu et al. \(2020\)](#) rely on MCMC to develop a driving cycle construction method for city buses in consideration of passenger load. These papers have shown that the MCMC approach can be useful in uncovering insights from large amounts of statistical data. However, one major drawback of the MCMC methods is that they require large amounts of data and do not produce consistent results ([Chen et al., 2022](#)).
4. Clustering analysis (CA) methods: For CA, oftentimes a dimensionality reduction technique such as principal component analysis (PCA) is combined with a clustering algorithm such as k-means, hierarchical clustering, or c-fuzzy clustering to classify micro-trips based on their component characteristics ([Peng et al., 2020](#); [Yang et al., 2020](#); [Fotouhi and Montazeri-Gh, 2013](#)). The dimensionality reduction allows the most salient features of trip segments to shine through, reducing the noise and/or redundancy in the data and allowing effective clustering to take place. CA enjoys strong theoretical support ([Chen et al., 2022](#)), and have been employed in many studies to resolve issues around real-world representativeness of driving cycles faced by RS. [Abas et al. \(2018\)](#) use a two-step clustering method to develop a driving cycle for urban conditions in Malaysia and show their energy consumption is within 10% error compared to the actual value. [Jing et al. \(2017\)](#) build a driving cycle for Tianjin based on two-class Fisher’s discrimination, a form of linear discriminant analysis, where the two classes are congestion and smooth traffic conditions. They show that it averages only 5.46% error compared to the real-world conditions. [Li et al. \(2019\)](#) also propose a two-level clustering method to determine the daily driving cycle for electric vehicles around China. The above studies demonstrate the effectiveness of CA in establishing statistical cycles. Often, determining the appropriate features to cluster on and the optimal number of clusters are the greatest challenges with these approaches. Although PCA or other dimensionality reduction techniques help ‘select’ the most influential or salient parameters, it diminishes the interpretability of the features on which clusters are based. This paper implements a version of CA that avoids the issues with parameter selection by clustering on time series shape.
5. Optimization-based (OB) methods: OB methods have been added to the cycle construction literature more recently than the others. Whereas the other categories are statistical, OB methods approach the creation of driving cycles as an optimization problem. Specifically, they view a driving cycle as an optimal combination of multiple smaller segments. From the smaller

segments, the optimization forms a driving cycle that best matches the target parameters from the real-world conditions seen in the original measured database. [Chen et al. \(2019\)](#) employ OB via genetic algorithm to construct a driving cycle for urban driving conditions in Shenyang, China. The constructed cycle is shown applied to the energy management system design for three vehicle types, and shown to be able to help significantly reduce energy cost under real driving conditions. [Cui et al. \(2021\)](#) use simulated annealing to develop a cycle in a case study in the Fujian Province of China, which better aligns speed-acceleration patterns with real-world driving characteristics over comparative cycles. [Cui et al. \(2022\)](#) develop driving cycles in Fuzhou city under various conditions with the min-max ant colony optimization algorithm. These methods tend to be more accurate than methods in the other categories but do not produce as consistent of results as CA, miss out on data transients, and requires large volumes of data to achieve the best accuracy.

2.3. Research gap

In the above methods, the constructed driving cycles are a representation of the mean parameters of the driver behavior seen in the measured database for the target cycle duration. Such statistical constructions can miss out on real world transient features seen in the measured database ([Ganesh Sundarkumar et al., 2021](#)). [Ganesh Sundarkumar et al. \(2021\)](#) address this issue by forming clusters of micro-trips based on their time-series shapes rather than their kinematic characteristics. Micro-trips are thus classified on the pattern of driving they exhibit, rather than mean features. Taking representative micro-trips from these clusters ensures that the real-world transients are preserved in the developed cycle.

However, previous time series shape based clustering methodologies are limited by their usage of the k-means algorithm, which cannot cluster micro-trips of varying length. This forces them to truncate all micro-trips to the same length, thus altering the pattern of driving they represent. Furthermore, previous approaches have not sequenced the candidate micro-trips from the final clusters in an order that represents that which would be seen on the road. This study seeks to address these gaps.

First, mixed integer programming (MIP) is utilized instead of k-means to cluster the micro-trips. The advantages of k-means are that it is relatively simple to implement, with quick compute time, and that its limitations are known and well studied ([Fränti and Sieranoja, 2019](#)). Typically, the main limitation of k-means is that it converges to locally optimal clustering solutions which are rarely the global optimum. This is because centroids cannot move between clusters if their distance is large, or if there are stable clusters in between that prevent centroid movement ([Fränti and Sieranoja, 2019](#)). However, k-means has another important limitation for time series shape-based clustering applications. In particular, it is not useful for minimizing ‘arbitrary’ non-Euclidean distances such as dynamic time warping (DTW) distances, which are used to measure the similarity between two time series of different lengths.

DTW works by finding an optimal alignment between two time series by stretching or compressing their respective temporal axes, allowing for the comparison of sequences with varying speeds or timing disparities. In general, clustering on DTW distances is the most effective way to cluster micro-trips by their shape, which is critical to capturing similar driving patterns. Longest common sub-sequence (LCSS) ([Vlachos et al., 2002](#)) and edit distance of real-number sequences (EDR) ([Chen et al., 2005](#)) are alternative methods for measuring similarity between time series or real number sequences. LCSS measures the similarity between two sequences by finding the longest sub-sequence that is common to both sequences. A sub-sequence is a sequence that can be derived from another sequence by deleting some or no elements. EDR measures the similarity between two sequences by calculating the minimum number of edit operations (insertions, deletions, substitutions) required to transform one sequence into the other. While LCSS and EDR are less computationally intensive than DTW, DTW is the best choice for constructing representative driving cycles that closely mimic the detailed driving patterns within micro-trips as it is well-suited for capturing fine-grained shape similarities and variations.

When attempting to cluster on arbitrary distances such as DTW distances, k-means may not converge or yield a good result, and it is therefore not suited to time series shape-based clustering for time series of different lengths. Previous research has circumvented this by truncating all micro-trips to the same length prior to clustering. However, altering the micro-trips in this manner distorts their shape, significantly affecting their representation of a driving pattern and the subsequent clustering results. Avoiding this distortion would let the benefits of time series shape based clustering shine through by allowing real-world transients to be completely preserved from the measured dataset.

Options to cluster micro-trips of varying lengths by DTW distance included MIP and density-based clustering algorithms such as DBSCAN or OPTICS. While the density-based algorithms can naturally identify and separate noisy outliers from the clusters, they require the minimum distance parameter ϵ . However, there may not be a straightforward mapping from ϵ values to meaningful time series distances in the case of arbitrary DTW distances. Furthermore, these algorithms may be sensitive to the initial conditions and converge to local optima.

MIP works by forming the clustering problem as an optimization problem, based on a matrix of pair-wise DTW distances between micro-trips. This means that in addition to successfully clustering micro-trips of different lengths, MIP also reliably finds the globally optimal clustering solution. In this study, to avoid including spam micro-trips of an arbitrarily short length in the clusters, a minimum micro-trip length parameter t_{\min} is enforced. Only micro-trips of duration greater than or equal to t_{\min} are extracted from the dataset.

Second, a simple maximum likelihood approach is employed to stitch micro-trips together in the final cycle. Previously, the candidate micro-trips selected from clusters had been stitched together in an arbitrary order. Using Viterbi programming to determine which order the clusters are most likely to present in the real world allows the final driving cycle to form a more accurate representation of trips seen in the measured database. Further details on the MIP clustering algorithm and the stitching approach are discussed in Section 3.

3. Methodology

This section details the proposed approach to driving cycle development. The steps in developing a driving cycle are route selection, data collection, cycle construction, and cycle assessment (Galgamuwa et al., 2015). In this study, after routes are selected and data is collected, cycle construction begins by smoothing the data with a rolling window average of 3s, and extracting micro-trips from the trip data. Dynamic time warping (DTW) is then used to calculate the pair-wise similarity between all micro-trips, and MIP is used to cluster them based on the DTW similarity. Next, candidate micro-trips are extracted from clusters. These candidate micro-trips are ordered based on a transition probability matrix between clusters, and then stitched together to form the final cycle. Finally, eight target parameters are used to evaluate the developed cycle against the original measured database to assess its performance value, or 'representativeness'. A flow chart for this proposed methodology is presented in Fig. 2. Each step is described in detail in the rest of this section.

The choice of target driving cycle duration varies widely in the literature. For example, driving cycles for buses in Delhi are only 170s (Badusha and Ghosh, 1999), but the Singapore Driving Cycle lasts 2400s (Ho et al., 2014). Results from Giraldo et al. (2021) indicate that cycles must last at least 1800s in order to reliably represent real-world driving patterns. However, longer cycles are more inconvenient to run in a laboratory setting. Therefore, target driving cycle duration in this study is restricted to being in the range of 1800–2400 s.

3.1. Route selection

Route selection is critical to the successful development of representative driving cycles. Historically, it has been a subjective process, where researchers attempt to use their knowledge of local traffic conditions, topography, road type, and other factors to select sample routes (Tong et al., 1999). More recent papers have been more systematic about route selection, employing techniques such as Annual Average Daily Traffic (AADT) (Hung et al., 2007) to locate the most frequently used routes. Other studies that deal with developing driving cycles for public transport have used other methodologies such as employing surveyors with an aim to cover different activity patterns at and across different urban districts and time periods (Tong and Ng, 2021a,b; Nesamani and Subramanian, 2011), or selecting existing routes a typical commuter or target vehicle would take regularly (Norbakyah et al., 2021; Kivekäs et al., 2018a; Tong and Ng, 2023; Liu et al., 2020). The dataset used here is captured on three frequently used routes of distinct type in and around Stellenbosch, South Africa, (urban, inter-urban, and hilly) and at three distinct times of day (morning, midday, evening) to capture variation in traffic conditions. Although the taxis take predominately the same path on each trip for each route, these routes are not fixed, and there are occasional slight deviations to drop off passengers in nearby neighborhoods. They are shown in Fig. 3. All routes are recorded to and from the Bergzicht taxi rank in Stellenbosch.

The urban route is from Stellenbosch to nearby township Kayamandi. The route is characterized by a speed limit of 60 km/h and bidirectional roads. The route varies slightly, since the taxis circle the neighborhood to drop passengers near their homes before terminating their trip at the rank. This is representative of typical taxi behavior in townships.

The inter-urban route is from Stellenbosch to Somerset West. It consists mostly of a dual carriageway with a maximum speed limit of 100 km/h.

The hilly route is from Stellenbosch to Pniel. This route has a maximum speed limit of 80 km/h, and crosses the Helshoogte mountain pass for a steady climb of approximately 300 m in 7 km.

3.2. Data collection

There are two primary methods of data collection in the driving cycle literature for collecting on-road vehicle speed-time data: the chase car approach and the on-board device approach (Tong et al., 2011). In the chase car approach, a trained driver carefully follows a selected target vehicle along a pre-defined route in a vehicle equipped with GPS. The on-board device approach involves carrying a device, such as a GPS device, that records vehicle speed-time data on board a target vehicle.

The chase car approach is more commonly utilized as it carries a lower resource requirement. However, it is less accurate than the on-board device approach, which directly captures the speed-time data for the target vehicle. Furthermore, Galgamuwa et al. (2015) note in a review of driving cycle methodologies that the chase car method is limited in areas where the driving pattern is aggressive. This study uses data collected with the on-board device approach. Specifically, the data is 1 Hz GPS tracking data gathered from minibus taxi paratransit vehicles traveling in and around Stellenbosch in the Western Cape province in South Africa. In order to address inaccuracies in the GPS data, the GPS data underwent a filtering process before being recorded on the SD card. Only fully accurate signals were retained, while any incomplete or erroneous ones were excluded. After data cleaning, there were 62 trips used in cycle construction, with a total of 79,855 observed seconds of raw data. Trips were taken with over 40 different paratransit drivers, to capture variation in individual driving style. Further details on the dataset and collection protocol are available in this repository: <https://data.mendeley.com/datasets/xt69cnwh56/1> (Hull et al., 2022).

3.3. Micro-trip extraction

A micro-trip is a segment of the trip that starts and ends when the vehicle is stopped. Vehicles are considered stopped when speed is less than 2 km/h and $-0.10 \text{ m/s}^2 < \text{acceleration} < 0.10 \text{ m/s}^2$ (Zhao et al., 2020). Before processing, the data is smoothed

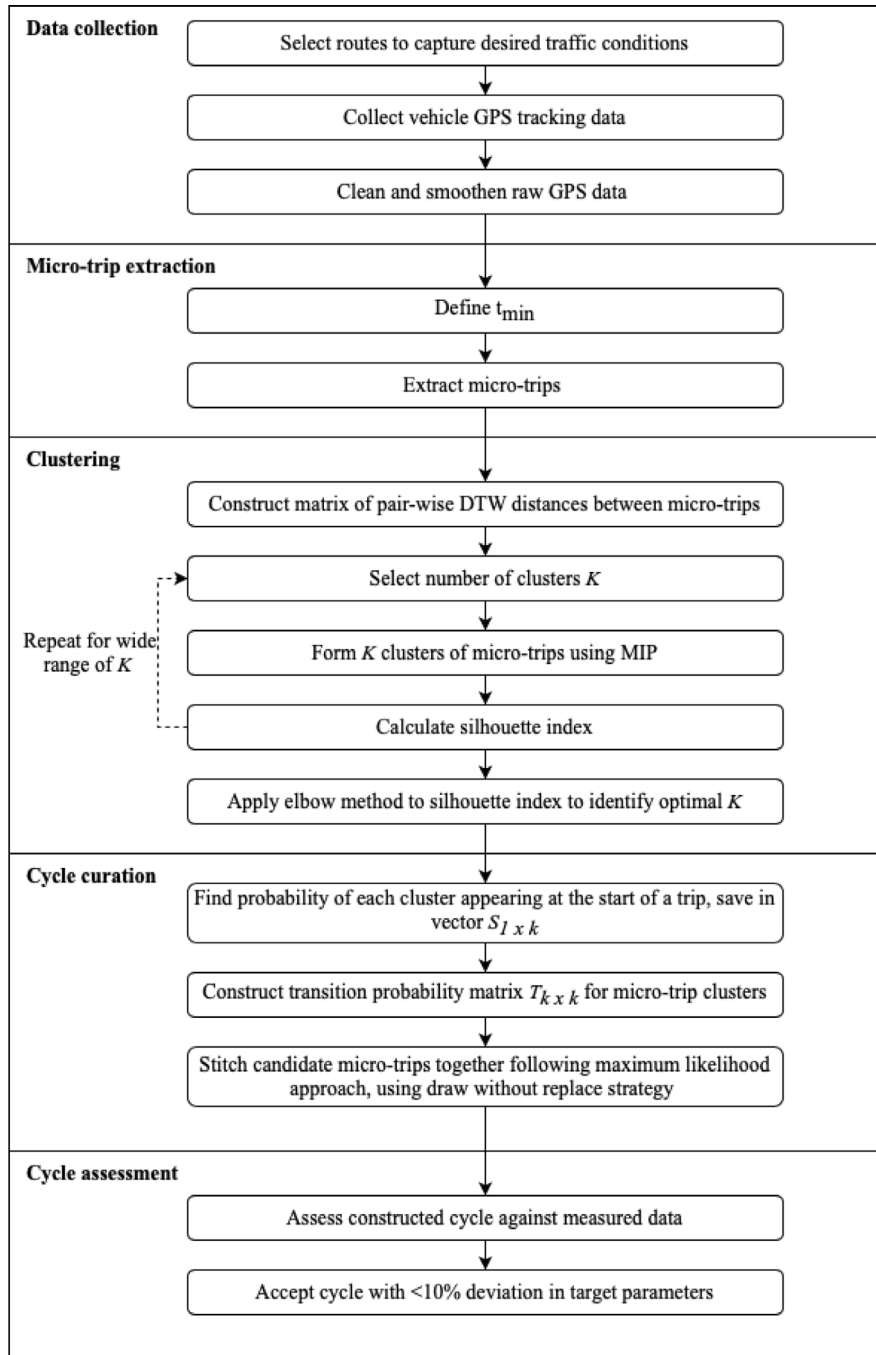


Fig. 2. Driving cycle construction and assessment methodology flow chart.

using a rolling window average of 3s, to eliminate outlier acceleration values from GPS noise. A rolling window average involves calculating the average value of a set of data points within a moving or ‘rolling’ window as it slides through the data from the beginning to the end. Windows of 4s and 5s were also considered, but 3s was sufficient to eliminate outlier acceleration values above 4s that may have been caused by vibration following the WLTP standard, and any widening the window any further would lead to a reduction in sensitivity to abrupt maneuvers, which is important in the context of paratransit driving.

The algorithm used to extract micro-trips from the measured database is presented in the following algorithm.

1. Read in trip file, index $0, \dots, i, \dots, T$.



Fig. 3. Bird's eye view of the three routes selected for data capture (Google, 0000).
Source: Figure extracted from Hull et al. (2023).

2. Search for the index of the first GPS sample observation where the vehicle is stopped. i must be greater than t_{\min} for $[0, i]$ to be a valid micro-trip.
3. Create a micro-trip by slicing trip file from index 0 to index i .
4. Save this slice in set of micro-trips.
5. If less than t_{\min} seconds remain in the trip after this slice (i.e. $T - i < t_{\min}$), append remainder of trip to previous micro-trip. This avoids the issues of either creating a micro-trip with length less than t_{\min} , or cutting off data at the end of trips. Else, reindex the remainder of the trip (so the index starts from 0 again) and repeat steps 2–4.
6. Repeat Steps 1–5 until all trips in measured database are exhausted.

In this study, micro-trips are restricted to segments lasting at least minimum duration t_{\min} . The parameter t_{\min} is enforced to ensure that each extracted micro-trip adequately represents a distinct pattern of driving, which will enable the clustering algorithm to perform better. Without t_{\min} , ‘spam’ micro-trips that do not represent a pattern of driving may be created. These spam micro-trips may then be erroneously clustered with micro-trips that do represent a distinct pattern of driving. Nevertheless, too large t_{\min} is also dangerous, since it may lead to the clumping of multiple shorter duration driving patterns in the same micro-trip. These shorter patterns can then be lost in the clustering stage. The ideal t_{\min} is the smallest possible value that avoids the creation of spam micro-trips.

Since ideal value for t_{\min} is unknown and likely context dependent, this study explores two values: $t_{\min} \in \{20, 180\}$. As stated previously, the smallest t_{\min} possible is preferable. Nevertheless, considering the 3-second rolling window average, a t_{\min} substantially shorter than 20 s would yield an insufficient number of data points to reliably capture consistent driving patterns. Conversely, it could require a more extended period, possibly even several minutes, to discern and establish a discernible driving pattern. Thus, the exploration of $t_{\min} = 180$ becomes warranted. Using $t_{\min} = 20$ yielded 931 micro-trips, $t_{\min} = 180$ yielded 232 micro-trips. Representative driving cycles are developed using both values and compared in Section 4.

Fig. 4 shows an example of how a trip is divided into micro-trips using $t_{\min} = 180$ s. As mentioned in the literature review, one of the contributions of this paper is to provide a methodology that can cluster micro-trip time series’ of varying lengths. This can be seen in Fig. 4, where the micro-trips vary in length from 180 s (the minimum length enforced by t_{\min} to over 200 s). The raw and smooth data are visualized, and since a rolling window average of only 3 s is used, the smooth data does not appear significantly different than the raw data, however outlier acceleration values are eliminated. The gray dashed line shows how the micro-trip does not end at the first stop, but at the first stop after t_{\min} is reached. The rationale supporting the desirability of a larger t_{\min} can be more clearly seen in the bottom chart in Fig. 4. Specifically, it is conceivable that the shorter segment spanning approximately 790 and 840s may not adequately represent a driving pattern. In this scenario, increasing the t_{\min} to a larger quantity such as 180s could enhance the performance of the clustering algorithm, making it more effective in grouping similar instances together and preventing the shorter trip from being treated as an isolated micro-trip.

3.4. Clustering

Clustering is the process of grouping a set of p data points into k clusters based on their similarity. The motivation for clustering in this application is to create groups of micro-trips that represent distinct driving patterns. Dynamic time warping (DTW) is used to compute the similarity of a pair of micro-trips. DTW calculates the similarity of micro-trips based on their shapes. The advantage of DTW over simple Euclidean distance is that it allows for the comparison of two series of different lengths, which Euclidean distance is unable to do effectively.

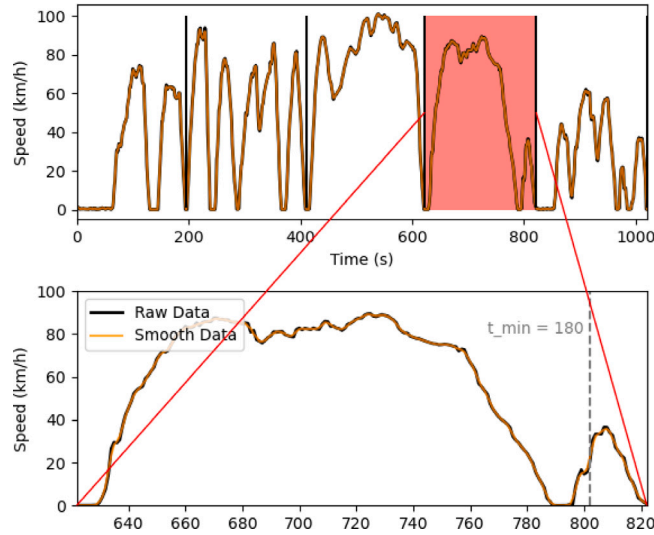


Fig. 4. Sample trip comparing raw (black) vs smooth (orange) data. The bottom plot zooms in on a micro-trip extracted from the trip extracted using $t_{\min} = 180$. The gray dashed line at 180 s shows how the micro-trip does not end at the first stop, but at the first stop after t_{\min} is reached.

Table 1
DTW constraints.

Constraint	Description
Boundary	x_1 and y_1 must match, and x_n and y_m must match. This guarantees that the entirety of both sequences is used.
Monotonicity	If $x_j > x_k$, and $y_l > y_i$, then the algorithm cannot match x_j with y_i and x_k with y_l . This guarantees that observations are not repeated.
Continuity	There can be no jumps in time. The mapping between X and Y must transition along consecutive points in time. This guarantees that no observations are left out.

At a high level, the clustering process can be understood in two steps. First, for a set of p micro-trips, a matrix of pair-wise DTW distances between every micro-trip is constructed, $D_{p \times p}$. Then, a set of K clusters are formed on the basis of the DTW distances using Mixed Integer Programming (MIP). The usage of MIP ensures that a globally optimal clustering solution is reached (Kumtepli et al., 2023).

Both the DTW distance calculations and MIP clustering are run with the DTW-C++ package (Kumtepli et al., 2023), which implements MIP with the Gurobi solver. The rest of this section describes how the DTW and MIP clustering are implemented in the DTW-C++ package, as well as how number of clusters k is chosen. The authors refer readers to Kumtepli et al. (2023) for further background on the mathematical implementation.

3.4.1. Dynamic time warping (DTW)

The formulation of DTW is detailed in the following algorithm, and the constraints are summarized in Table 1. Consider two time series, X and Y , of length n and m respectively:

$$X = (x_1, x_2, \dots, x_n)$$

$$Y = (y_1, y_2, \dots, y_m)$$

1. Iterate along X , computing the Euclidean distance from each point in X to each point in Y .
2. Store the minimum distance calculated.
3. Switch to using Y as a reference series and repeat previous steps.
4. Add up minimum distances stored with objective of minimizing cumulative total distance, subject to the constraints in Table 1.

This cumulative sum W is the DTW distance between the two time series.

To summarize the constraints in Table 1: the beginning and end of the mapping between the two time series must match in position, there must be no cross-matches, and all points from both series must be included in the mapping.

The DTW distance W is calculated between every micro-trip in the set of micro-trips. Each DTW distance is stored in $D_{p \times p}$, where element $d_{i,j}$ gives the distance between micro-trips i and j . Integer programming can then use the matrix D to split the time series' into k separate clusters.

	j_1	j_2	j_3	j_4	j_5
i_1	0	0	0	0	0
i_2	1	1	0	0	1
i_3	0	0	0	0	0
i_4	0	0	1	1	0
i_5	0	0	0	0	0

A

Fig. 5. Example matrix A . An entry of 1 indicates that micro-trip j belongs to cluster with centroid i . Cluster centers are non-zero rows (highlighted), where each cluster member is 1 and non-members are 0.

Source: From Kumtepe et al. (2023)

3.4.2. Mixed integer programming (MIP) clustering

The following describes how to split the matrix $D_{p \times p}$ into k clusters using MIP.

A binary square matrix $A_{p \times p}$ is constructed, where $A_{ij} = 1$ if micro-trip j is a member of the i th cluster centroid and 0 otherwise, as shown in Fig. 5. Centroids are represented by non-zero diagonal entries in A .

The optimization problem is then given by Eq. (1):

$$A^* = \min_A \sum_i \sum_j D_{ij} \times A_{ij} \quad (1)$$

The optimization is subject to the constraints in Eqs. (2), (3), and (4).

1. Only k series can be centroids

$$\sum_{i=1}^p A_{ii} = k \quad (2)$$

2. Each micro-trip time series can only exist in one cluster

$$\sum_{i=1}^p A_{ij} = 1 \quad \forall j \in P \quad (3)$$

with $P = \{1, 2, \dots, p\}$.

3. A row can only contain non-zero entries if the corresponding diagonal is non-zero, so a micro-trip time series can only be in a cluster where the row corresponds to a centroid micro-trip.

$$A_{ij} \leq A_{ii} \quad \forall i, j \in P \quad (4)$$

After solving this MIP problem, the cluster centroids are represented by the non-zero diagonal entries of A , and the members of each cluster are given by the corresponding non-zero columns in A .

In Fig. 5, the clusters are 1,2,5 and 3,4 with 2 and 4 being the cluster centroids.

3.4.3. Choosing number of clusters

Choosing the number of clusters (k) is recognized as the most difficult part of most clustering problems. If k is too high, data of the same type will be split into multiple clusters (overclustering). If k is too low, some patterns in the data may be missed (underclustering).

The elbow method, as traditionally applied, is primarily used to determine the optimal number of clusters based on the Within-Cluster Sum of Squares (WCSS). It involves plotting the WCSS against the number of clusters and looking for an “elbow” point in the curve to identify the optimal cluster count. It aims to strike a balance between simplicity (few clusters) and accuracy (capturing meaningful patterns in the data).

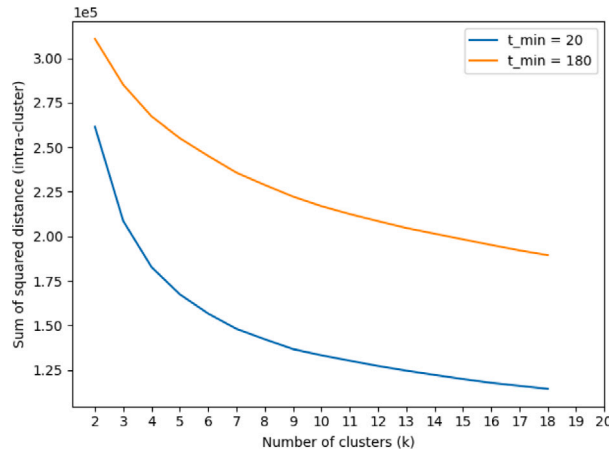


Fig. 6. Within cluster sum of squares (WCSS) scores for clusters formed with $t_{\min} = 20$ and $t_{\min} = 180$.

WCSS represents the sum of squared distances between data points and their assigned cluster centers, or in other words, how ‘tight’ the clustering is. A smaller WCSS indicates that the data points are closer to their cluster centers. To obtain WCSS, one must compute how far each data point X_j in the cluster is from the centroid. For the i th cluster C_i with centroid C_i^{centroid} containing n data points, WCSS is given by Eq. (5).

$$WCSS = \sum_{i=1}^K \sum_{j=1}^n (X_j - C_i^{\text{centroid}})^2 \quad (5)$$

The elbow method is informed by where there is an ‘elbow’ in the graph – a place where the rate of decrease starts to diminish more rapidly.

This point is considered the optimal number of clusters because it represents a balance between having enough clusters to capture meaningful patterns in the data and avoiding over-segmentation. Because the set of micro-trips with a higher t_{\min} has fewer observations, it is logical that it would contain a lower number of distinct driving patterns and therefore a lower optimal choice for k . This expectation is borne out in Fig. 6, which indicates the optimal k is eight for the set of micro-trips extracted with $t_{\min} = 20$, and six for the set extracted with $t_{\min} = 180$.

3.5. Candidate micro-trip selection

After clustering, candidate micro-trips are selected for usage in the final cycle. Following previous clustering-based methods in the literature, they are selected by taking micro-trips near to the cluster centroids until target driving cycle duration is reached (Ganesh Sundarkumar et al., 2021; Fotouhi and Montazeri-Gh, 2013; Zhao et al., 2020; Peng et al., 2020). In this case, target duration is between 1800 and 2400 s. Although the mean trip length in the data is just above 20 min, in real-world driving, routes and conditions can vary significantly, and not every trip will follow the same pattern or duration. Giraldo et al. (2021) find that cycles must last at least 1800 s in order to reliably represent real-world driving patterns. For example, in this study, as described in Section 3.1, the data is collected on three distinct route types (urban, inter-urban, and hilly) at three distinct times of day (morning, midday, and evening). This variability is important for testing and evaluating vehicles and systems under different conditions. A longer driving cycle captures a broader range of driving scenarios and provides a more comprehensive representation of real-world driving behavior. However, Giraldo et al. (2021) find that cycles that are too long become inconvenient to run in a laboratory setting, so 2400 s is selected as an upper bound, being twice as long as our mean trip length and equal in duration to the longest driving cycle in previous literature (Ho et al., 2014).

The number of candidate micro-trips from each cluster is determined by cluster weight C_w . C_w is the size of the cluster proportional to that of the smallest cluster, rounded to the nearest integer (Eq. (6)).

$$C_w = \text{round}\left(\frac{\text{len}(C)}{\text{len}(C^{\min})}\right) \quad (6)$$

where C^{\min} is the cluster containing the fewest micro-trips. The logic behind this approach is that the frequency of a cluster’s appearance in the representative driving cycle should be proportional to the frequency of its appearances in real world data.

3.5.1. Stitching the driving cycle

After the candidate micro-trips are selected, they must be stitched together to form the representative driving cycle. This study takes a maximum likelihood approach for sequencing candidate micro-trips in the final cycle.

A transition probability matrix $T_{k \times k}$ is created, where each element $T_{w,z}$ represents the probability that a micro-trip from cluster w in the original database is followed by a micro-trip from cluster z . Additionally, the likelihood of each cluster appearing at the start of a trip is computed and added to a vector $S_{1 \times k}$. This probability is determined for a cluster by dividing the number of times a constituent micro-trip appears at the beginning of a trip by the total number of trips.

Once the transition probability matrix T and the initial probability vector (S) are established, they are used to assemble a representative driving cycle. The assembly process follows a “draw without replacement” strategy, commencing with the cluster that has the highest probability of being the starting point of a trip, as indicated by cluster with the maximum value in S .

Maximum transition probabilities from T are utilized to determine the next cluster to include in the sequence. In other words, the next cluster is chosen based on the cluster with the highest likelihood of being followed when transitioning from the current cluster.

In summary, this procedure creates a representative driving cycle by modeling the probabilities of transitioning between clusters of micro-trips and considering the likelihood of each cluster appearing at the beginning of a trip. The goal is to generate a realistic sequence of clusters based on observed patterns in the original trip database. The procedure is outlined in the following algorithm.

Procedure: Stitching Representative Driving Cycle (RDC) from candidate micro-trips

1. Initialize Representative Driving Cycle (*RDC*) as an empty list.
2. Start with the cluster that has the highest probability of being seen at a trip's beginning by setting $C_{curr} = \text{argmax } S$.
3. Append the candidate micro-trips from C_{curr} to *RDC*.
4. Look up the index of C_{curr} in matrix T to find C_{next} .
5. Remove C_{curr} from T .
6. $C_{curr} = C_{next}$
7. Repeat steps 3–6 until T is empty. When T is empty it means all clusters are now present in the *RDC*.

Key:

RDC: Representative Driving Cycle (list)
 C_{curr} : Current cluster being considered (cluster)
 C_{next} : Next cluster to be considered (cluster)
 S : Initial probability vector (vector)
 T : Transition probability matrix between clusters (matrix)

3.6. Cycle assessment

For cycle assessment, characteristic parameters of the constructed driving cycle are compared against the measured real world driving data. The relative error (RE_i) of characteristic parameter (x_i) in the driving cycle is the percentage difference from the value of that parameter in the measured database (x_i^*).

$$RE_i = \frac{x_i^* - x_i}{x_i^*} \quad (7)$$

Performance value (PV) for a cycle is the mean of the absolute values of the RE across all characteristic parameters.

$$PV = \frac{1}{N} \sum_{i=0}^N |RE_i| \quad (8)$$

where N is the number of characteristic parameters used for cycle assessment.

Many characteristic parameters have been considered for usage for driving cycle assessment in the literature. These parameters fall into six categories: distance, duration, speed, acceleration/deceleration, dynamics, and stop data. Berzi et al. (2016) provide a list 40 parameters from these six categories that are used in the literature. However, it is not beneficial to use an excessive number of parameters. Some parameters are highly correlated, and if too many are included, then certain characteristics will be overemphasized in the final driving cycle (Zhao et al., 2020; Arun et al., 2017; Pouresmaeili et al., 2018). Consequently, most research uses a subset of parameters. This study chooses eight popular parameters in the literature that have been found to have a significant effect on driving pattern and little correlation between them (Chen et al., 2022; Zhao et al., 2020). The parameters and the expressions used to calculate them are listed in Table 2.

4. Results

The RE of each characteristic parameter and final PV for the developed cycles are shown in Table 3. The cycle developed with $t_{min} = 20$ had the better PV of 3.65%, versus 7.14% for the cycle developed with $t_{min} = 180$. The cycles are plotted in Fig. 7. The cycle data series' are available in this repository (link forthcoming).

The less restrictive value of t_{min} allowed for the development of a more representative driving cycle, i.e. one which had lower average deviance from the measured dataset across the eight parameters. This result indicates the existence of distinct driving patterns of a duration between 20 and 180s. Using $t_{min} = 180$ means clustering algorithm is unable to distinguish these patterns from others.

This result rings intuitively true in the context of paratransit. The nature of paratransit driver behavior is such that there are many aggressive maneuvers that last a short period of time (Zeeman and Booyesen, 2014). Micro-trips that last significantly longer than these maneuvers would miss out on isolating these unique driving patterns.

Table 2
Characteristic parameters (*CP*) used for cycle assessment.

<i>CP</i>	Unit	Expression
Average speed	km/h	$V = \frac{\sum_{i=1}^n v_i}{n}$
Average running speed	km/h	$V_r = \frac{\sum_{i=1}^n v_i}{n} \forall [V > 2]$
Average acceleration	m/s ²	$a = \frac{\sum_{i=1}^n a_i}{n} \forall [a > 0.10]$
Average deceleration	m/s ²	$d = \frac{\sum_{i=1}^n a_i}{n} \forall [a < -0.10]$
Standard deviation acceleration	m/s ²	$\sigma_a = \sqrt{\frac{\sum_{i=1}^n (a_i - \bar{a})^2}{n}}$
% time accelerating	%	$P_a = \frac{t_{acc}}{t_{total}} \forall [a > 0.10]$
% time decelerating	%	$P_d = \frac{t_{dec}}{t_{total}} \forall [a < -0.10]$
% time cruising	%	$P_{cr} = \frac{t_{cru}}{t_{total}} \forall [-0.10 \leq a \leq 0.10]$

Table 3

Characteristic parameter (*CP*) values for the measured dataset and developed representative driving cycles, and performance values for the latter.

<i>CP</i>	Measured data	$t_{\min} = 20$		$t_{\min} = 180$	
		Value	Relative error (%)	Value	Relative error (%)
Duration (s)	–	2077	–	2151	–
V (km/h)	35.30	35.87	1.62	34.03	–3.59
V_r (km/h)	43.18	44.19	1.29	44.29	2.56
a (m/s ²)	0.58	0.59	2.01	0.66	14.18
d (m/s ²)	–0.66	–0.71	–7.54	–0.70	–7.53
σ_a (m/s ²)	0.66	0.67	1.29	0.70	6.35
P_a (%)	37.79	36.69	–2.92	33.74	–10.72
P_d (%)	33.07	30.27	–8.49	31.30	–5.34
P_{cr} (%)	14.74	15.20	3.02	13.74	–6.85
PV (%)	–	–	3.65	–	7.14

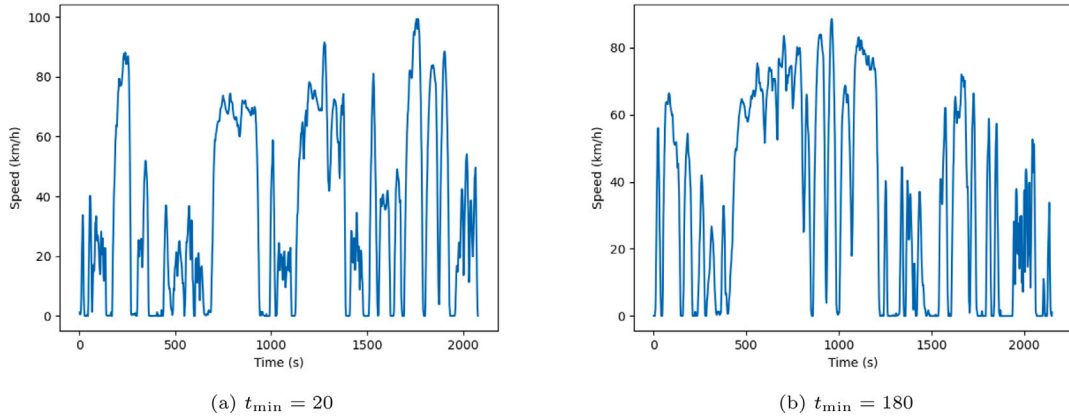


Fig. 7. Developed representative driving cycles using $t_{\min} = 20$ s (a) and $t_{\min} = 180$ s (b).

4.1. Comparison with international driving cycles

A comparison between several international driving cycles and the paratransit driving cycle (PDC) developed in this paper shows that the PDC represents the real world operating conditions of paratransit most accurately. The comparison of the characteristic parameters from the Worldwide Harmonized Light Vehicles Test Cycle (WLTC) (DieselNet, 2019.01), Federal Test Procedure 75 (FTP-75) (DieselNet, 2014.08), China Light-Duty Vehicle Test Cycle (CLTC) (DieselNet, 2020.12), the PDC, and the measured data is given in Table 4. Compared to the performance value (PV) of 3.65% for the PDC, the PV s of the WLTC, FTP-75, and CLTC are 33.87%, 23.36%, 25.82% respectively, indicating that the PDC is a better fit for the real-world conditions faced by paratransit vehicles. Furthermore, unlike the other three, the method used to construct the PDC ensures it contains transients from data captured on paratransit vehicles. These transients are unique to paratransit driving behavior, are difficult to replicate through any sort of statistical construction, and improve its representativeness of the true vehicle movements.

Notably, the average acceleration (a) and deceleration (d), as well as the percentage of time spent accelerating (P_a) and deceleration (P_d) are substantially greater for the PDC and measured data than for any of the three comparison cycles, while percentage of time spent cruising P_{cr} is lower. This is indicative of aggressive driving behavior of paratransit vehicles, which are known to frequently engage in rapid acceleration/deceleration maneuvers.

Table 4

Comparison of the characteristic parameters and performance value of the paratransit driving cycle (PDC) developed in this study versus several international driving cycles.

CP	Data	PDC		WLTC		FTP-75		CLTC	
		Val.	RE (%)	Val.	RE (%)	Val.	RE (%)	Val.	RE(%)
V (km/h)	35.30	35.87	1.62	46.50	31.73	33.89	−3.99	28.96	−17.69
V_r (km/h)	43.18	44.19	1.29	53.15	23.09	25.82	−40.20	37.15	−13.96
a (m/s ²)	0.58	0.59	2.01	0.42	−27.59	0.51	−12.07	0.45	−22.41
d (m/s ²)	−0.66	−0.71	−7.54	−0.44	33.33	−0.58	12.12	−0.49	25.76
σ_a (m/s ²)	0.66	0.67	1.29	0.43	−34.85	0.56	−15.15	0.48	−27.27
P_a (%)	37.79	36.69	−2.92	30.90	−18.23	31.1	−17.70	28.6	−24.32
P_d (%)	33.07	30.27	−8.49	28.60	−13.52	27.1	−18.05	26.4	−20.17
P_{cr} (%)	14.74	15.20	3.02	27.8	88.60	24.7	67.57	22.8	54.68
PV (%)	–	–	3.65	–	33.87	–	23.36	–	25.82

5. Conclusion

This work develops a representative driving cycle for paratransit vehicles traveling around Stellenbosch, South Africa, and proposes a novel time series shape based clustering methodology for cycle development. Dynamic time warping is combined with mixed integer programming to cluster micro-trips of varying lengths based on their time series shapes. Formulating the micro-trip clustering problem as an optimization problem enables the globally optimal clustering solution to be found. Micro-trips nearest to cluster centroids were used in the final representative driving cycle. The utilization of the original micro-trips in the cycle construction process ensures that diverse real world transients from the real-world driving data are present in the representative cycle, which capture micro-level mobility movements that are otherwise difficult to capture using other driving cycle construction methodologies.

The concept of t_{\min} , which defines minimum acceptable micro-trip length, is explored to modulate the driving patterns captured by the clustering algorithm. Cycles constructed with varying t_{\min} are evaluated. The shorter duration t_{\min} (20 s) is found to lead to a more representative cycle than the longer duration t_{\min} (180 s) in this instance, but it is possible that greater t_{\min} would be more suitable in other contexts. Future work is needed to understand and compare the contexts that lend themselves to shorter or longer minimum acceptable micro-trip length.

The representative driving cycle developed in this study, the Paratransit Driving Cycle (PDC), is found to be less than 10% deviant from the measured database in each of eight characteristic parameters, with an average deviance or “performance value” of 3.65%. This is a significant improvement over several international comparison cycles, namely the WLTP, FTP-75, and CLTC, which exhibit performance values of 23%–34%. Furthermore, compared to these cycles, the PDC expresses high average acceleration and deceleration and high percentage of time spent accelerating and decelerating, and low percentage of time cruising. This is characteristic of paratransit driving behavior, which is commonly identified as aggressive and known to engage in relatively frequent acceleration/deceleration maneuvers. The PDC speed-time data series is available in this Mendeley repository (link forthcoming).

The method proposed in this study is flexible and can be adapted for various contexts to construct representative driving cycles in other cities and countries, which can then be used to determine the typical operating conditions that paratransit vehicles face in these contexts. As the global transportation system evolves and electrifies, representative cycles will be important for establishing an in-depth understanding of vehicle operational needs, efficiently framing transport system policies, and guiding long term vehicle design and marketing with respect to decisions such as engine power, fuel type (e.g., diesel or electric), and vehicle size. The cycles and methodology developed here serve as a platform for understanding paratransit mobility, but further investigations into cycle characteristics such as stop time distributions are required to garner more actionable insights from a transportation system planning perspective.

The mobility style of paratransit, although not well studied, is unique and unconventional, requires more specific analysis in various contexts across sub-Saharan Africa. This work is a step in this direction, but additional work is needed to gather data and establish driving cycles for paratransit vehicles in urban and peri-urban areas in the region.

CRedit authorship contribution statement

Christopher Hull: Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft. **Katherine A. Collett:** Methodology, Supervision, Writing – review & editing. **Malcolm D. McCulloch:** Conceptualization, Methodology, Resources, Writing – review & editing.

Data availability

A link to the data is provided in the manuscript in the Data collection section.

References

- Abas, M.A., Rajoo, S., Abidin, S.F.Z., 2018. *Transp. Res. D* 63, 388–403.
- Abraham, C.J., Rix, A., Booysen, M.J., 2023. Aligned simulation models for simulating Africa's electric minibus taxis. <http://dx.doi.org/10.36227/techrxiv.23376311.v1>, TechRxiv, TechRxiv. Preprint.
- Abraham, A.J., Ndiabaty, I., Booysen, M., 2021. *Energy Sustain. Devel.* 64, 118–127.
- Arun, N., Mahesh, S., Ramadurai, G., Nagendra, S.S., 2017. *Sustain. Cities Soc.* 32, 508–512.
- Axsen, J., Plötz, P., Wolinetz, M., 2020. *Nature Clim. Change* 10 (9), 809–818.
- Badusha, A.A., Ghosh, B., 1999. Technical Report, SAE Technical Paper.
- Behrens, R., McCormick, D., Mfinanga, D., 2015. *Paratransit in African Cities: Operations, Regulation and Reform*, first ed. Routledge, London, pp. 1–311.
- Berzi, L., Delogo, M., Pierini, M., 2016. *Transp. Res. D* 47, 299–322.
- Bishop, J.D., Axon, C.J., McCulloch, M.D., 2012. *Transp. Res. D* 17 (5), 389–397.
- Booyesen, M.J., Abra8468964ham, C.J., Rix, A.J., Ndiabaty, I., 2022. *Energy Res. Soc. Sci.* 85, 102403.
- Brady, J., O'Mahony, M., 2016. *Appl. Energy* 177, 165–178.
- Chandrashekar, C., Agrawal, P., Chatterjee, P., Pawar, D.S., 2021. *IATSS Res.* 45 (4), 551–560.
- Chen, Z., Fang, Z., Zhang, Q., Zhou, N., Yu, Q., 2022. *Trans. Inst. Measur. Control* 01423312221094384.
- Chen, L., Özsü, M.T., Oria, V., 2005. *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*. pp. 491–502.
- Chen, Z., Zhang, Q., Lu, J., Bi, J., 2019. *Energy* 186, 115766.
- Collett, K.A., Hirmer, S.A., 2021. *Nat. Sustain.* 4 (7), 562–564.
- Cui, Y., Xu, H., Zou, F., Chen, Z., Gong, K., 2021. *Energy* 235, 121434.
- Cui, Y., Zou, F., Xu, H., Chen, Z., Gong, K., 2022. *Energy* 247, 123455.
- Dai, Z., Niemeier, D., Eisinger, D., 2008. *Driving Cycles: A New Cycle-Building Method That Better Represents Real-World Emissions*, Vol. 570. Department of Civil and Environmental Engineering, University of California, Davis.
- Degrauwe, B., Weiss, M., 2017. *Environ. Pollut.* 222, 234–241.
- DieselNet, 2014.08. FTP-75.
- DieselNet, 2019.01. Worldwide Harmonized Light Vehicles Test Cycle (WLTC).
- DieselNet, 2020.12. China Light-Duty Vehicle Test Cycle (CLTC).
- Dixit, M., Sivakumar, A., 2020. *Transp. Res. D* 87, 102473.
- Estevés-Booth, A., Muneer, T., Kirby, H., Kubie, J., Hunter, J., 2001. *Transp. Res. D* 6 (3), 209–220.
- Evans, J., O'Brien, J., Ch Ng, B., 2018. *Trans. Inst. British Geogr.* 43 (4), 674–688.
- Fotouhi, A., Montazeri-Gh, M., 2013. *Sci. Iranica* 20 (2), 286–293.
- Fränti, P., Sieranoja, S., 2019. *Pattern Recognit.* 93, 95–112.
- Galgamuwa, U., Perera, L., Bandara, S., 2015. *J. Transport. Technol.* 05, 191–203.
- Ganesh Sundarkumar, G., B. V., S.B., Munigety, C.R., Arora, A.S., 2021. *Transp. Res. D* 97, 102896.
- Giliomee, J., Hull, C., Collett, K.A., McCulloch, M., Booysen, M., 2023. *Transp. Res. D* 118, 103728.
- Giraldo, M., Quirama, L.F., Huertas, J.I., Tibaquirá, J.E., 2021. *World Electr. Veh. J.* 12 (4), 212.
- Gong, H., Zou, Y., Yang, Q., Fan, J., Sun, F., Goehlich, D., 2018. *Energy* 150, 901–912.
- Google, ND. 0000. Google Earth, URL <https://earth.google.com/web/@-33.88614835,18.82391936,130.63246104a,12436.34033288d,35y,350.20430081h,0t,0r>.
- Gupta, D., Garg, A., 2020. *Transp. Res. D* 85, 102474.
- Ho, S.-H., Wong, Y.-D., Chang, V.W.-C., 2014. *Atmos. Environ.* 97, 353–362.
- Hull, C., Giliomee, J., Collett, K.A., McCulloch, M.D., Booysen, M.J., 2022. 1Hz GPS Tracking Data on Minibus Taxi Paratransit Vehicles in South Africa. Mendeley Data.
- Hull, C., Giliomee, J., Collett, K.A., McCulloch, M.D., Booysen, M., 2023. *Transp. Res. D* 118, 103695.
- Hung, W.T., Tong, H., Lee, C., Ha, K., Pao, L., 2007. *Transp. Res. D* 12 (2), 115–128.
- Jaikumar, R., Nagendra, S.S., Sivanandan, R., 2017. *Transp. Res. D* 54, 397–409.
- Jing, Z., Wang, G., Zhang, S., Qiu, C., 2017. *Transp. Res. D* 53, 78–87.
- Kamble, S.H., Mathew, T.V., Sharma, G.K., 2009. *Transp. Res. D* 14 (2), 132–140.
- Kaymaz, H., Korkmaz, H., Erdal, H., 2019. *Transp. Res. D* 75, 123–135.
- Kim, M., Kim, H.K., 2020. *Transp. Res. D* 85, 102464.
- Kivekäs, K., Vepsäläinen, J., Tammi, K., 2018a. *IEEE Access* 6, 55586–55598.
- Kivekäs, K., Vepsäläinen, J., Tammi, K., 2018b. *IEEE Access* 6, 55586–55598.
- Kumtepel, V., Perriment, R., Howey, D.A., 2023. Fast dynamic time warping and clustering in C++. arXiv preprint arXiv:2307.04904.
- Li, Y., Peng, J., He, H., Xie, S., 2017. *Energy Procedia* 105, 3219–3224.
- Li, X., Zhang, Q., Peng, Z., Wang, A., Wang, W., 2019. *J. Clean. Prod.* 206, 827–837.
- Liu, X., Ma, J., Zhao, X., Du, J., Xiong, Y., 2020. *J. Adv. Transport.* 2020, 1–21.
- Mayakuntla, S.K., Verma, A., 2018. *Transp. Res. D* 65, 725–735.
- McCormick, D., Schalekamp, H., Mfinanga, D., 2016. In: Behrens, R., McCormick, D., Mfinanga, D. (Eds.), *Paratransit in African Cities: Operations, Regulation and Reform*, first ed. Routledge, New York, pp. 59–78.
- Michel, A., 2004. *Sci. Total Environ.* 334, 73–84.
- Ndiabaty, I., Booysen, M.J., 2020a. *J. Transp. Geogr.* 88.
- Ndiabaty, I., Booysen, M., 2020b. In: Miriam, C., Paul, C. (Eds.), *IST-Africa 2020 Conference Proceedings*. IST-Africa Institute and IIMC, Kampala, Uganda, pp. 1–10.
- Ndiabaty, I., Booysen, M.J., 2021. *J. Transp. Geogr.* 92, 103001.
- Nesamani, K., Subramanian, K., 2011. *Atmos. Environ.* 45 (31), 5469–5476.
- Norbakyah, J., Nordiyana, M., Anida, I., Ayob, A., Salisa, A., 2021. *Int. J. Electr. Comput. Eng.* 11, 2054.
- Nylund, N., Erkkilä, K., Clark, N., Rideout, G., 2007. Evaluation of duty cycles for heavy-duty urban vehicles : Final report of IEA AMF Annex 29.
- Peng, Y., Zhuang, Y., Yang, Y., 2020. *Proc. Inst. Mech. Eng. D* 234 (2–3), 714–724.
- Pfrieem, M., Gauterin, F., 2016. *World Electr. Veh. J.* 8 (1), 14–24.
- Pouresmaeil, M.A., Aghayan, I., Taghizadeh, S.A., 2018. *Sustain. Cities Soc.* 36, 12–20.
- Quirama, L.F., Giraldo, M., Huertas, J.I., Jaller, M., 2020. *Transp. Res. D* 82, 102294.
- Rodríguez, R.A., Virguez, E.A., Rodríguez, P.A., Behrentz, E., 2016. *Transp. Res. D* 43, 192–206.
- Saleh, W., Kumar, R., Sharma, A., 2010. *World J. Sci. Technol. Sustain. Develop.*
- Shen, P., Zhao, Z., Li, J., Zhan, X., 2018. *Transp. Res. D* 59, 346–360.
- Smith, R., Shahidinejad, S., Blair, D., Bibeau, E., 2011. *Transp. Res. D* 16 (3), 218–224.
- Tong, H., 2019. *Sustainable Cities Soc.* 48, 101588.

- Tong, H., Hung, W., Cheung, C.S., 1999. *Atmos. Environ.* 33 (15), 2323–2335.
- Tong, H.Y., Ng, K.W., 2021a. *Environ. Sci. Pollut. Res.* 28, 14343–14357.
- Tong, H.Y., Ng, K., 2021b. *Sustainable Cities Soc.* 69, 102854.
- Tong, H.Y., Ng, K.W., 2023. *Sustain. Cities Soc.* 98, 104819.
- Tong, H., Tung, H., Hung, W., Nguyen, H., 2011. *Atmos. Environ.* 45 (29), 5191–5199.
- Toyota, 0000. **Toyota HiAce specifications.**
- Toyota, 2023. **Toyota Hi'Ace Ses'fikile.**
- Tsai, J.-H., Chiang, H.-L., Hsu, Y.-C., Peng, B.-J., Hung, R.-F., 2005. *Atmos. Environ.* 39 (35), 6631–6641.
- Vámosi, A., Czégé, L., Kocsis, I., 2022. *Period. Polytech. Transport. Eng.* 50 (2), 184–190.
- Vlachos, M., Kollios, G., Gunopulos, D., 2002. *Proceedings 18th International Conference on Data Engineering. IEEE*, pp. 673–684.
- Wang, Q., Huo, H., He, K., Yao, Z., Zhang, Q., 2008. *Transp. Res. D* 13 (5), 289–297.
- Wangsness, P.B., Proost, S., Rødseth, K.L., 2020. *Transp. Res. D* 86, 102384.
- Yang, Y., Li, T., Zhang, T., Yu, Q., 2020. *Sustain. Cities Soc.* 53, 101949.
- Zeeman, A.S., Booysen, M., 2014. *Southern African Transport Conference.*
- Zhao, X., Yu, Q., Ma, J., Wu, Y., Yu, M., Ye, Y., 2018. *J. Adv. Transp.* 2018, 1–18.
- Zhao, X., Zhao, X., Yu, Q., Ye, Y., Yu, M., 2020. *Transp. Res. D* 81, 102279.