# Collecting population-representative bike-riding GPS data to understand bike-riding activity and patterns using smartphones and Bluetooth beacons

Debjit Bhowmick[1*], Danyang Dai[1], Meead Saberi[2], Trisalyn Nelson[3], Mark Stevenson[4], Sachith Seneviratne[4], Kerry Nice[4], Christopher Pettit[5], Hai L. Vu[6] and Ben Beck[1]

[1*]School of Public Health and Preventive Medicine, Monash University, 553 St. Kilda Road, Melbourne, 3004, VIC, Australia.
[2]School of Civil and Environmental Engineering, Research Centre for Integrated Transport Innovation (rCITI), University of New South Wales, Sydney, Australia.
[3]Department of Geography, University of California, Santa Barbara, USA.
[4]Transport, Health and Urban Systems Research Lab, Melbourne School of Design, The University of Melbourne, Melbourne, Australia.
[5]School of Built Environment, University of New South Wales, Sydney, Australia.
[6]Institute of Transport Studies, Monash University, Melbourne, VIC, Australia.

*Corresponding author(s). E-mail(s): debjit.bhowmick@monash.edu;

**Abstract**

Bike-riding GPS data offers detailed insights and individual-level mobility information which are critical for understanding bike-riding travel behaviour, enhancing transportation safety and equity, and developing models to estimate bike route choice and volumes at high spatio-temporal resolution. Yet, large-scale bicycling-specific GPS data collection studies are infrequent, with many existing studies lacking robust spatial and/or temporal coverage, or have been influenced by sampling biases leading to these data lacking representativeness. Additionally, accurately detecting bike-riding trips from continuously collected raw GPS data without human intervention remains a challenge. This study presents a novel GPS data collection approach by leveraging the combination of a smartphone application with a Bluetooth beacon attached to a participant's bike. Aided by minimal heuristic post-processing, our method limits data collection to trips taken by bike without the need for participant intervention, carefully optimising between survey participation, privacy challenges, participant workload, and robust bike-riding trip detection. Our method is applied to collect 19,782 bike trips from 673 adults spanning eight months and three seasons in Greater Melbourne, Australia. The collected dataset is shown to represent the underlying adult bike-riding population in terms of demographics (sex, occupation and employment type), temporal and spatial patterns. The average trip length (median = 4.8 km), duration (median = 20.9 min), and frequency of bicycling trips (median = 2.7 trips/week) were greater among men, middle-aged and older adults. The 'Interested but Concerned' riders (classified using Geller typology) rode more frequently, while the 'Strong and Fearless' and 'Enthused and Confident' groups rode greater distances and for longer. Participants rode on roads/streets without bike infrastructure for more than half of their trips by distance, while spending 24% and 17% on off-road paths and bike lanes respectively. This population-representative dataset will be key in the context of urban planning and policymaking.

**Keywords:** Cycling, GPS data collection, Active transport, Mobility data analysis

# 1 Introduction

## 1.1 Background

Policy-makers are looking to promote the uptake of bike-riding as a healthy mode of travel [1–3] that reduces the negative effects of traditional motorised transport such as physical inactivity, air pollution, and traffic congestion, and achieves sustainability goals. However, fears about riding alongside motor vehicle traffic and the lack of safe and appropriate bike-riding infrastructure are significant barriers [4, 5]. For the strategic installation of safer bike-riding infrastructure and the implementation of pro-bicycling policies in general, rigorous evidence-informed scientific studies is necessary, which in turn rely on high-quality bicycling data, which is scarce [6]. Bicycling-specific GPS data can reveal valuable insights on actual individual-level bike-riding behaviour, as well as help understand overall bike-riding trends and activity patterns in a geographical area. Such data also contributes towards the development of robust bike-riding route choice and volume models at high spatial resolution [7–9] which are key in the sustainable urban planning context.

GPS data collection involves greater complexity of recruitment, needing far greater levels of participant engagement to collect data, beyond just a short survey. Therefore, high-quality bicycling GPS data with appropriate spatial and temporal coverage is scarce. Additionally, bicycling-specific GPS data collection has its own set of challenges. There is a growing utility of continuously-collected GPS data using smartphones [7, 10, 11] as they enable the collection of comprehensive mobility-related information from a user, with reduced participant workload, and bypassing recall errors [12, 13]. However, such continuous data collection methods also return large quantities of redundant data such as when the participant is at home or workplace for large periods of time [14, 15], consume significant amounts of smartphone battery and data uploads [16, 17], and can develop privacy concerns in users [18] leading to a reduced willingness to participate and an increased rate of study dropouts [13, 19]. Furthermore, accurately detecting only bicycling trips or bicycling trip segments (in the case of multi-modal trips) from continuously collected GPS data is challenging [15, 20–22]. To bypass these shortcomings, continuous data collection is often substituted by ‘phased sampling’ approaches, where GPS data is only collected during trips made by the user, thereby conserving smartphone energy and data [14]. However, significant challenges remain related to trade-offs between resource efficiency and data accuracy [14]. Manual intervention approaches are often utilised, such as the user reporting the trip start and end times themselves [23]. However, this introduces biases and errors in data collection, such as under-reporting of trips due to self-reporting bias (when done in real-time), loss of contextual information [19], or recall error (when trip validation is done post data collection period) [13]. While Bluetooth beacons have been proven to provide micro-location where traditional location services have limited access [24, 25], they have been rarely implemented to collect large-scale GPS data [26].

Crowdsourced bicycling data mitigates some limitations of standalone GPS data [7] by providing processed GPS data at a high spatial and temporal resolution covering large areas, thus offsetting limitations related to spatial coverage of traditional datasets, and real-time monitoring of mobility [27, 28]. Strava, a popular social workout app, collects self-reported GPS data from its users, and this dataset is popular among researchers [6, 7, 28–31]. However, the demographic features of Strava users (and other crowdsourced data platforms in general) are skewed due to self-selection bias [23], particularly towards males and people aged 25-44 years [32], and towards recreational cycling as compared to utilitarian cycling [28]. Furthermore, while crowdsourcing applications such as Strava collect GPS data from their users, they only offer aggregated information available for download, such as aggregate count information in each link in the network, or aggregated counts of trips across different origins and destination sectors along with temporal information. [28]. Hence, raw GPS data collected directly from individual bicyclists gives access to more disaggregated information at the individual-level and trip-level, such as greater semantic information such as the origin and destination of a trip, start and end time, trip duration, travel speeds, chosen route, and socio-demographic information of the rider (in specific cases)[15].

Furthermore, the mobility behaviour of cyclists differs significantly from that of motor vehicle users. Cycling behaviour is critically dependent on available infrastructure, safety and safety perception, weather and other disaggregated factors, which are not significant drivers of mobility behaviour of motor vehicle users [33]. Additionally, there are locally-specific contextual factors such as utilitarian cycling culture [34, 35], socio-economics [36], programs, policies [35, 37] and legislation [38] that influence cycling patterns significantly. Therefore, bicycling route choices and behaviour, in general, are more complex and are found to have significant disparity spatially, across countries, and often across different cities in the same country. In comparison, for example, motor vehicle drivers will usually opt for the fastest routes to their destinations irrespective of their city or country [39]. Given the highly localised behaviour of bicyclists, bicycling GPS datasets are less transferable, and therefore, there is a need for dedicated bicycling GPS datasets for individual study areas, and often for individual studies.

## 1.2 State of research

The improved feasibility of large-scale GPS data collection due to the ubiquity of smartphones and recent developments in location-based services [11, 23, 40–43] has led to the advent of a host of bicycling-specific GPS data collection studies in recent times. Such studies have taken place predominantly in developed nations. In North America, studies have been conducted in South Minneapolis [44], Oregon [10, 41], Los Angeles [40], California [11, 45, 46], Texas [47], Ohio [48], and Montreal [42]. In Europe, similar data collection exercises have taken place in a host of cities including Zurich [49], Copenhagen [50], Dresden [15, 23], Noord-Brabant [51], Madrid [43], Gdynia [52], Bologna [53, 54], Oslo [55], and Amsterdam [56]. Systematic reviews by Pritchard [57] and Łukawska [58] mention a more exhaustive list of other bicycling-specific GPS data collection studies. A notable example of a nationwide multi-modal GPS data collection is the MOBIS project in Switzerland [59]. Across Australia, bicycling GPS data was collected using the Riderlog app developed by Bicycle Network (an Australian cycling membership and advocacy organisation) and provided the platform for studies conducted in Sydney [60, 61]. However, Riderlog does not collect data anymore as it is no longer supported, and the existing dataset is outdated given significant changes in infrastructure and bicycle ridership across the major cities in Australia.

Existing bicycling GPS data collection studies (see Table 1) either continuously collect GPS data or rely on phased sampling approaches, which either require extensive preprocessing or contain biases due to excessive participant intervention. Most of the existing bicycling GPS datasets have limited spatial [49, 53] and temporal coverage [40, 53], except for the ones that were collected as part of some continental or national data collection initiatives, but are now discontinued and outdated (RiderLog GPS data). Also, most bicycling GPS data collection studies did not report a deliberate attempt at collecting a population-representative sample, nor did they report any statistical comparisons against population-level distributions (distribution of the population across classes of any relevant demographic attribute such as gender, age, employment status) with the distributions in their sample (distribution of the sample across the same demographic classes). Lißner et. al. (2021) recommended the application of population-level weights derived from household travel or mobility surveys to GPS datasets to generate population-representative samples [23]. Population-representative GPS data samples are more likely to generate population-representative results and assist in the development of population-representative and calibrated models. Therefore, representative datasets are key in the context of urban planning. This is critical in the case of bicycling (and not so much for motorised modes) as bicycling travel behaviour is governed significantly by the socio-demographic (age, gender) and spatial characteristics (place of residence, access to safe bicycling infrastructure) of the participant. Furthermore, to the best of our knowledge, no study except one [53] has reported the spatial representativeness of their collected GPS data. Therefore, to overcome the limitations of crowdsourced data and mitigate the challenges of continuous GPS data collection and participant intervention to self-record bike trips, there is a need for adopting a bicycling GPS data collection approach having a large spatial and temporal coverage that optimises the trade-offs between survey participation, participant workload, and accurate and relevant bicycling GPS data collection, while robustly capturing individual-level bicycling patterns and behaviour across a diversity of population sub-groups. Such a GPS data collection approach is absent from the literature. However, opportunities lie to leverage learnings from other applications to advance our ability to collect more representative bicycling GPS data at higher spatial resolutions. The benefits of collecting a population-representative bicycling GPS dataset lie in the ability to make robust and reliable interpretations in future studies that hold true for the underlying bicycling population in the study area.

**Table 1**: Existing bicycling-specific GPS data collection exercises (not systematically reviewed)

| **Authors/Institution (Data collection years)** | **Study area** | **Size of study area (sq km)** | **Duration of data collection** | **Individuals** | **Trips** | **Data collection method** | **Demographic *(Spatial)* representativeness** |
|---|---|---|---|---|---|---|---|
| Dutch Cyclists' Union, National Bike Counting Week Fietselweek (2015) | The Netherlands | 41,865 | 1 week | 38,000 | 377,321 | Not documented | Not documented |
| Bella Mossa initiative (2017) [54] | Bologna, Italy | 141 | 6 months | | 270,000 | Not documented | Not documented |
| RiderLog GPS data, Bicycle Network (2010-2013) | 8 Greater Capital City Statistical Areas, Australia | | 3.5 years | 7,601 | 120,085 | Not documented | Not documented |
| M. Łukawska, M. Paulsen, T.K. Rasmussen, A.F. Jensen, and O.A. Nielsen (2019-2021) [50] | Copenhagen, Denmark | 2,778 | 20 months | 6,523 | 134,169 | Phased-sampling. Users had to switch on the Bluetooth in their helmet to start recording GPS data. | Not documented |
| G. Menghini, N. Carrasco, N. Schussler and K. W. Axhausen (2004) [49] | Zurich, Switzlerland | 88 | | 2,435 | 73,493 | Not documented | Not documented |
| F. Rupi, C. Poliziani and J. Schweizer (2016) [53] | Bologna, Italy | 141 | 1 month | 1,123 | 27,348 | Not documented | Reported representative gender balance. *(Significant correlation between GPS data and traditional counts.)* |
| J. Strauss, L. F. Miranda-Moreno and P. Morency (2013) [42] | Montreal, Québec, Canada | | 5 months | 1,000 | 10,000 | Phased-sampling. Users manually recorded their trips. | Not documented |

**Table 1**: Existing bicycling-specific GPS data collection exercises (not systematically reviewed)

| **Authors/Institution (Data collection years)** | **Study area** | **Size of study area (sq km)** | **Duration of data collection** | **Individuals** | **Trips** | **Data collection method** | **Demographic *(Spatial)* representativeness** |
|---|---|---|---|---|---|---|---|
| G. Romanillos and M. Zaltz Austwick (2013-2014) [43] | Madrid, Spain | 604 | 16 months | 328 | 6,022 | Not documented | Not documented |
| B. Charlton, E. Sall, M. Schwartz and J. Hood (2009-2010) [11] | San Francisco, California, USA | 600 | 5 months | 952 | 5,178 | Phased-sampling. Users manually recorded their trips. | Reported oversampling of men, frequent cyclists. Comparisons of other variables not reported. No statistical comparisons were reported. |
| S. Lißner and S. Huber (2018) [23] | Dresden, Germany | 329 | 4 months | 187 | 4,951 | Phased-sampling. Users manually recorded their trips. | Mentions that their sample is representative. While it does report descriptive details of study participants, it does not report any statistical comparisons with population-representative datasets. |
| J. G. Hudson, J. C. Duthie, Y. K. Rathod, K. A. Larsen and J. L. Meyer (2011) [47] | Austin, Texas, USA | 846 | 6 months | 317 | 3,198 | Phased-sampling. Users manually recorded their trips. | Reported similar gender distribution compared to a 2002 survey, oversampling of "expert bicyclists". No statistical comparisons were reported. |
| Y. Park and G. Akar (2016) [48] | Columbus, Ohio, USA | 586 | 3 months | 78 | 1,531 | Phased-sampling. Users manually recorded their trips. | Not documented |

**Table 1**: Existing bicycling-specific GPS data collection exercises (not systematically reviewed)

| **Authors/Institution (Data collection years)** | **Study area** | **Size of study area (sq km)** | **Duration of data collection** | **Individuals** | **Trips** | **Data collection method** | **Demographic *(Spatial)* representativeness** |
|---|---|---|---|---|---|---|---|
| J. Broach, J. Dill and J. Gliebe (2007) [41] | Portland, Oregon, USA | 375 | 9 months | 154 | 1,449 | Not documented | Reported oversampling of women, and people who were older, more educated, and full-time workers. Comparisons of other variables not reported. No statistical comparisons were reported. |
| H. Francis and K. Krizek (2006) [62] | South Minneapolis, Minnesota, USA | | 2 months | 51 | 852 | Continuous data collection. | Not documented |
| P. Chen, Q. Shen and S. Childress (2009-2014) [46] | Seattle, Washington State, USA | 368 | 3.5 years | 197 | 544 | Phased-sampling. Users manually recorded their trips. | Not documented |
| S. Reddy, K. Shilton, G. Denisov, C. Cenizal, D. Estrin and M. Srivastava (2010) [40] | Los Angeles, California, USA | | 2 weeks | 12 | 208 | Phased-sampling. Users manually recorded their trips. | Not documented |

## 1.3 Aim

The study aims to demonstrate the feasibility of collecting bicycling GPS data from a sample representative of the underlying adult bike-riding population across a large spatial area (Greater Melbourne). We assess the feasibility of a novel bicycling GPS data collection system that allows for automatic capture of bike trips with minimal participant workload. We also propose an approach for capturing a population-representative sample and enabling quantification of representativeness.

# 2 Methods

## 2.1 Data collection

To achieve the stated aims, we set up a data collection method consisting of the following steps.

(a) Obtain existing population-representative datasets such as from subsets of household travel survey data with the mode cycling
(b) Set up the pre-data collection questionnaire based on the population-representative datasets to enable comparisons such as collecting relevant demographic variables
(c) Set up an appropriate sampling strategy to obtain a population-representative sample
(d) Recruit participants as per the sampling strategy and collect data
(e) Compare the study sample with the chosen population-representative datasets to demonstrate feasibility

### 2.1.1 Study area

We conducted a prospective observational study of bicycle trips taken by adults (18 years and older) in the Greater Melbourne area. Greater Melbourne is one of the Greater Capital City Statistical Areas (GCCSAs) (geographical areas that are designed to represent the functional extent of each of the eight state and territory capital cities) of Australia. In June 2018, Greater Melbourne covered an area of 9986 square kilometres, with 4.96 million residents. As per the Victorian Integrated Survey of Travel and Activity (VISTA) 2012-2020 data, bicycling mode share is a mere 1.8% on weekdays and 1.4% on weekends. 85% of the available road network that allows bicycling, does not have any bicycling infrastructure, with a little over 10% being shared or dedicated bike paths, and approximately 3% having either an associated protected, painted, or advisory bike lane [63]. The remaining 2% of the network was classified as other types of bike infrastructure.

### 2.1.2 Sampling strategy

Our original aim was to select a study sample that is representative of the adult bike-riding population in Greater Melbourne using a proportional stratified sampling approach [64]. Leveraging population-level household travel survey data (VISTA) [65], population-representative survey data [66] and urban bike-riding typologies [67], we developed strata based on age, gender, urban area and interest in bike-riding (as defined by the Geller typology; excluding non-riders who are defined as "no way no how") [68]. Geller typologies are important in representative sampling because there is a need for the sample to be representative not just in terms of bicyclist demographics but also bicycling travel behaviour. We chose the Geller typology for this study due to its ability to inform policy and practice, and the fact that it allowed for comparisons to prior studies that have used the same questions. For details on the Geller typology questions, please refer to [66]. The aim was then to apply the proportional stratified sampling approach which involves taking random samples from stratified groups, in proportion to the population to maximise the representativeness of the sample. However, due to slightly lower-than-expected participant numbers, we could not execute this sampling approach completely and thus the sample included in this study reflects a convenience sample (a form of non-probability sampling method where survey participants are selected for inclusion in the sample because they are the most convenient for the researcher to access). Nonetheless, we provide comparisons of our study sample to the broader bike-riding population of Greater Melbourne to quantify the representativeness of the study sample (further information is provided in Section 2.2.5).

### 2.1.3 Recruitment and survey design

We recruited participants via multiple channels, including key project stakeholders such as Bicycle Network, VicHealth, Parents' Voice, the Amy Gillett Foundation, WeRide Australia, the Municipal Association of Victoria, local councils, Bicycle User Groups (BUGs), and social media channels. Participants were recruited on a rolling basis meaning they were recruited at different times across a period spanning six months. Participants were eligible to participate in the study if they:

- completed the survey including their contact details,
- owned a bicycle, and
- had ridden their bike within the past 12 months.

Participants consented to the collection of relevant socio-demographic and mobility information via a survey. Consequently, they collected smartphone GPS data (including location coordinates, timestamps, and speeds) for two months individually. Due to the rolling recruitment structure, participants started data collection at different times of the year providing us bicycling GPS data across a diverse range of seasons. Participants were recruited throughout the data collection period that started in January 2022 and was completed in August 2022, covering summer, autumn (fall) and winter seasons. The survey captured information on the socio-demographics of the participants (age, gender, income, occupation, employment status, primary language, bike ownership, type of bike owned), their mobility behaviour (main mode of transport, frequency of bike rides, purpose of bike rides) and a set of questions to categorise participants according to their comfort riding in different street and path environments (the Geller typology) [68].

To overcome the limitations of prior approaches, we developed and employed a method to capture all bicycling trips that did not rely on participants having to manually log trips, but rather utilised a smartphone application that automatically captured bicycling trips with high accuracy. To achieve this, participants were mailed a Bluetooth beacon to be attached to their bicycles and were asked to download a smartphone application 'Ethica', which ran continuously in the background. Ethica [A] is an end-to-end research platform that enables researchers to quantitatively measure human behaviour using smartphones, wearables, and big data. Once installed, the Ethica app connected with their Bluetooth beacon when it was in range and only then did it start collecting high-frequency GPS data (collected at 1Hz), therefore ensuring a greater likelihood of capturing movement-related data only in instances when the participant is riding their bicycle. The Bluetooth-pairing feature was the advantage of using Ethica to collect GPS data. However, our approach is easily transferable by using any smartphone application that has similar capabilities. At the end of their two-month data collection period, participants were instructed to return the Bluetooth beacon to be eligible to participate in a prize draw involving e-bikes, bike-related memberships and vouchers.

Shared bike rides do not fall within the scope of this study. Shared bike rides are often limited by a geographic area, in the case of Melbourne, in which shared schemes are only available in inner Melbourne. Hence, the travel behaviour of shared bike users is distinct from non-shared bike users in terms of origins, destinations, demographics, and often, route choice. Furthermore, the latest release of VISTA does not include, or at least distinguish shared bike users, and therefore, comparisons with population-level estimates would not be possible.

## 2.2 Data processing

For answering specific research questions related to the mobility of bicyclists, raw GPS data requires multiple levels of preprocessing. We processed our data across multiple steps as follows, ass illustrated in Figure 1:

### 2.2.1 Noise filtering

*Noise filtering* involves filtering out erroneous and noisy GPS data points [69, 70]. Raw GPS data tends to be noisy and sometimes imprecise, especially when in indoor and semi-indoor situations (such as in a bus or train), and outdoors in urban canyons. First, we filtered out points when the location and timestamp of two consecutive points resulted in a speed greater than 100 km/h [23]. Second, we only kept data points that were collected either via GPS satellites or via nearby cell towers and Wi-Fi access points. Third, we

[A] https://ethicadata.com/

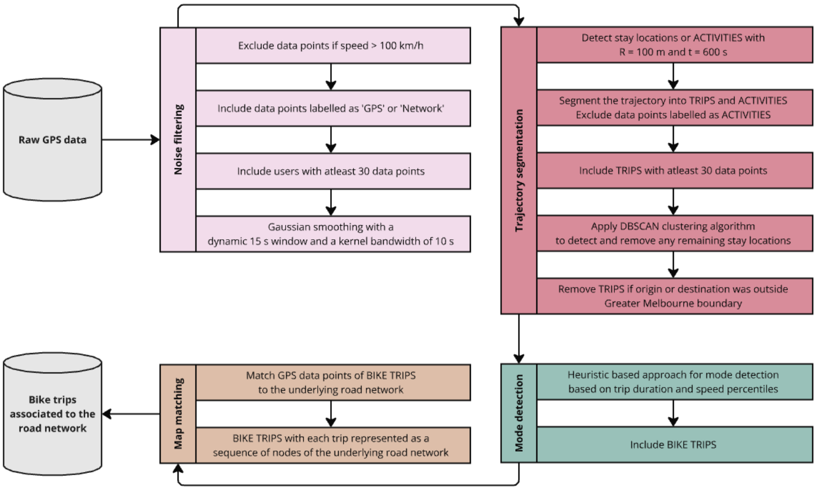



**Figure 1**: GPS data processing steps

excluded any user from any further preprocessing and analysis if their entire GPS dataset contained less than 30 data points, roughly corresponding to 30 seconds of data if collected continuously [50], as it is highly unlikely that a bike trip can be observed within that timeframe. It must be noted that we did not filter out any GPS data points based on their reported accuracy values. Fourth, we implemented the *Gaussian smoothing* function with a dynamic 15-second window and a kernel bandwidth of 10 seconds to smooth out the erratic raw speed gradients [15].

### 2.2.2 Trajectory segmentation

To partition the continuous GPS trajectory data stream into meaningful segments namely ‘trips’ and ‘activities’, and to remove the activity segments, we implemented a heuristic-based trajectory segmentation algorithm, leveraged from previous studies [23, 69, 71], which involved the concepts of temporal thresholds and spatial clustering techniques [23, 72–74]. We tuned the critical parameters to adapt to localised mobility behaviour in Melbourne, such as search radius and dwell time, detected gaps signifying the end of trips in the data stream, identified GPS points related to activities, and hence distinguished trips from activities. Corresponding details are available in Appendix A.

### 2.2.3 Mode detection

The novel approach employed in this study enabled automatic capture of bike trips. However, there were a limited number of scenarios in which non-bike trips could have been detected. For example, when a participant took their bike on a train as part of a multi-modal trip, or when a participant took their bike on or in their car. In both of these situations, the app would have collected continuous GPS data due to proximity between the smartphone and the beacon. To deal with these situations and any other erroneous data collection, we developed and employed a mode detection algorithm to remove non-bike trips. Typically, most mode detection algorithms do not have the capability to detect bike trips accurately, or their results for bicycle mode detection are poor relative to other modes [15, 75–78]. However, our objective was primarily to remove non-bike trips, given the high recall of our method. Thus we employed a heuristic-based algorithm (refer to Appendix B) to filter out non-bike trips similar to [15]. We removed trips with low-frequency data collection and then used thresholds of speed percentiles, average speed, speed differences and trip duration to remove non-bike trips. Finally, we removed any bike trips that were less than two minutes, as there was a high likelihood that these trips were incorrectly classified as bike trips. The threshold of two minutes was derived from Ethica's data collection method, where the Ethica application on the user's smartphone checks for the Bluetooth beacon every two minutes.

### 2.2.4 Map-matching

After obtaining individual bike trips, we map matched GPS points to an appropriate subset of the underlying road network, the bicycling network of Greater Melbourne to determine the most likely route taken by the bike rider [79]. By associating a route (a sequence of road segments) with a trip and a user, it was possible to determine the corresponding road network-related information. For our study, we have used a map-matching package coded in Python known as *Leuven.MapMatching*, proposed by [80], which is also used by multiple map-matching service providers, such as Valhalla, Mapbox, and GraphHopper [81]. First, we downsampled our high-frequency GPS data from a sampling rate of 1 second to 15 seconds to optimise time complexity and the completeness and accuracy of map-matching. We compared the results on a generous sample of our trajectories, original versus revised sampling rate. The results were not significantly different in terms of completeness and accuracy, as was indicated by [82–84]. Second, we downloaded a road network using OpenStreetMap (www.openstreetmap.org) [85] and composed a graph using a Python package known as OSMnx [86]. We chose to download road network data corresponding to a single time point, $30^{th}$ April 2022, the midpoint of our data collection period, while acknowledging that the underlying bicycling network might have undergone changes. We attempted to accommodate all streets and paths that could be possibly availed by a bike rider, excluding freeways and footpaths exclusive to pedestrians. Details of OpenStreetMap tags and values used for extracting the bicycling network graph can be found in Appendix E. Third, we map-matched all the bike trips on this graph using Leuven.MapMatching's *DistanceMatcher* class. We successfully map matched over 98% (19474 out of 19782) of bike trips.

### 2.2.5 Comparing study sample with population-level data

To gauge the representativeness of our survey sample, we compared our survey sample with bike-riding population-level estimates of Greater Melbourne. For demographics such as age, gender, occupation and employment status, we derived the population-level estimates of bike riders across Greater Melbourne from the Victorian Integrated Survey of Travel and Activity (VISTA) 2012-2020 data. This household travel survey is conducted throughout the year across Greater Melbourne, and other key regional centres periodically to understand average daily travel behaviour. Randomly selected households are asked to collect their travel data for a single specified day. VISTA employs a stratified, clustered sampling methodology, with stratification based on Local Government Areas (LGAs). Clusters were based upon the Mesh Block, the smallest unit within the Australian Statistical Geographical Standard. The survey and resulting data are then weighted to generate adult population-representative data at the LGA level. Our estimates correspond to the most recent release of VISTA data, i.e. from 2012-2020. VISTA does not report population estimates of Geller typologies. Therefore, for comparison of our survey participants with population-level estimates of Geller typology, we have referred to [66] where a survey was conducted on a representative sample of 3523 adults across Greater Melbourne and had classified them into one of the four Geller typologies.

We further conducted statistical tests for comparisons of trip characteristics and their distributions between samples. We conducted Mann-Whitney U-test to compare the difference of continuous trip metrics such as trip distance and duration between two independent groups such as trips made by men versus women (test statistic $U$), and a Chi-square goodness-of-fit test (test statistic $\chi^2$) to compare the proportion of counts (distribution) inside a class (gender) across each attribute level (male, female) with the corresponding population-level distribution.

### 2.2.6 Understanding the spatial representativeness of our dataset

To infer the spatial representativeness of our GPS dataset, we first divided our study area into SA2s (Statistical Area 2 defined by the Australian Bureau of Statistics) [87]. Then, we referred the study by Beck et. al. [67], which had developed an urban biking typology, grouping all SA2s having similar typologies across Greater Melbourne into five distinct clusters. We calculated the population-level proportion of bike trip origins across each of the five clusters using VISTA 2012-2020 data. Then, we calculated the distribution of bike trip origins across the same five clusters in our GPS data sample. Finally, we performed a Chi-squared goodness-of-fit test to determine statistically significant spatial representativeness. The Chi-squared test makes statistical comparisons between the frequency distribution of a categorical variable of two samples, which in this case are, the sample-level vs the population-level bike trip origin count distribution across the five clusters.

### 2.2.7 Understanding the usage of bike infrastructure

After map-matching the bike-riding GPS trajectories to the underlying Greater Melbourne bicycling road network that was classified based on a combination of bike infrastructure and functional class of the road [63, 88], we were able to generate insights on what infrastructure types were chosen by the survey participants using information from 19474 bike trips. The classes included Arterial Road – Mixed Traffic, Arterial Road - Painted Bike Lane, Collector Road - Mixed Traffic, Collector Road - Painted Bike Lane, Local Road - Mixed Traffic/Sharrow, Local Road - Painted Bike Lane, Protected Bike Lane, Off-road Bike Path, and Other. Mixed Traffic indicates road segments devoid of any type of bike infrastructure. Painted Bike Lane indicates on-road bike lanes that are separated from motorised traffic by a solid white painted line with the lane painted green on occasion. Protected Bike Lane indicates on-road bike lanes that are physically separated and thus protected from motorised traffic via a physical barrier. Sharrows indicate streets without a specific bicycle lane but with painted arrows and bicycle symbols indicating priority to cyclists. Off-road Bike Path indicates off-road paths that are either dedicated to cyclists or shared among pedestrians and cyclists.

# 3 Results

## 3.1 Size of final survey dataset

We initially recruited 903 participants who completed the screening survey and were subsequently sent a Bluetooth beacon at their preferred address. After executing the trajectory segmentation algorithm, 33,630 meaningful trip segments were identified. After mode detection, 21,640 trips were identified as bike trips, of which 1858 were removed as they were below 2 minutes. This left us with 19,782 bike trips (corresponding to 35.6 million GPS points) collected by 673 adult bike riders from Greater Melbourne, making it a significantly large standalone bicycling GPS data collection exercise placed after 6 studies in Table 1. Each of the 673 participants had completed at least one bike trip. In the following sections, we describe the characteristics of these 673 participants and their corresponding 19,782 bike trips.

## 3.2 Description of survey participants

Nearly half of our participants were aged between 35-54 years at the time of recruitment (49.3%). Participants who identified as female made up one-third of the participants (32.8%), while two-thirds identified as male (66.1%). The majority of participants were classified as ‘Interested but concerned’ according to the Geller typology (83.5%), while a further 16.2% were classified as either ‘Strong and fearless’ or ‘Enthused and confident’. Given our survey had a small proportion of ‘Strong and fearless’ participants (2.4%), we merged this category with the ‘Enthused and confident’ (13.8%) and reclassify them as ‘Strong and fearless’ or ‘Enthused and confident’, a typology representing cyclists who have significantly greater confidence in riding with traffic on roads. In terms of occupation, more than half of the participants identified as ‘Professionals’ (57.5%), while a further 17.1% were ‘Managers’. In terms of car usage, 17.7% of participants used a car daily while most participants used it at least once a week but not daily (56.5%), and only 5.3% stated that they never used a car in the last 12 months. In terms of bike trip frequencies, 26.9% of participants rode a bike daily, while a further 68.6% rode a bike at least once a week but not daily. Most participants used a conventional pedal bike (with no electric assist) (88.6%), while 9.9% people used an e-bike, and 1.5% owned both. These results have been tabulated in Table 2.

## 3.3 Comparison with population-level data

To understand the bias of our sample relative to the population of current people who ride, we have statistically and graphically compared the socio-demographic attributes of our participants with respective population-level numbers. Statistically, only the distribution of sex was not significantly different between our sample and adult population-level estimates of bike-riding ($\chi^2 = 0.006$, $p = 0.93$). For age, the difference was significant ($\chi^2 = 187.9$, $p \leq 0.01$), as we under-sampled younger and over-sampled older age groups, details of which are shown in Table 2. For Geller typology ($\chi^2 = 83.5$, $p \leq 0.01$), employment status ($\chi^2 = 32.3$, $p \leq 0.01$) and occupation ($\chi^2 = 0.80$, $p = 0.85$), the differences were statistically significant. In addition to the statistical comparisons, we have graphically illustrated the distributions of demographic characteristics of survey participants against corresponding adult bike-riding population-level proportions in Figure 2 and Figure 3. Slight differences were observed in Geller typology, as we oversampled the ‘Strong and fearless’ and ‘Enthused and confident’ categories. For the type of employment, the distributions were fairly similar with oversampling of full-time workers over casual workers. As for occupation, the sample-level and population-level distributions were fairly comparable for the majority of the occupation categories, albeit with some sampling biases.

## 3.4 Trip characteristics

We report trip characteristics of participants with detailed figures in Table 3 while Figure 4a and 4b show the travel time distributions and Figure 5a and 5b show the number of trips distribution of our sample compared to corresponding population-level estimates from VISTA data. It must be noted that we did not present the same plots for travel distance as the distance reported in VISTA uses the simulated shortest path distance, instead of an actual route distance. It can be observed in Table 3 that the number of trips across age, gender and Geller typology showed a distribution that was similar to the distribution of the underlying sample in terms of participant numbers. More than half of the bike trips (51.7%) were recorded by people aged between 35 years and 54 years old. Weekly bike trip frequency was the highest among

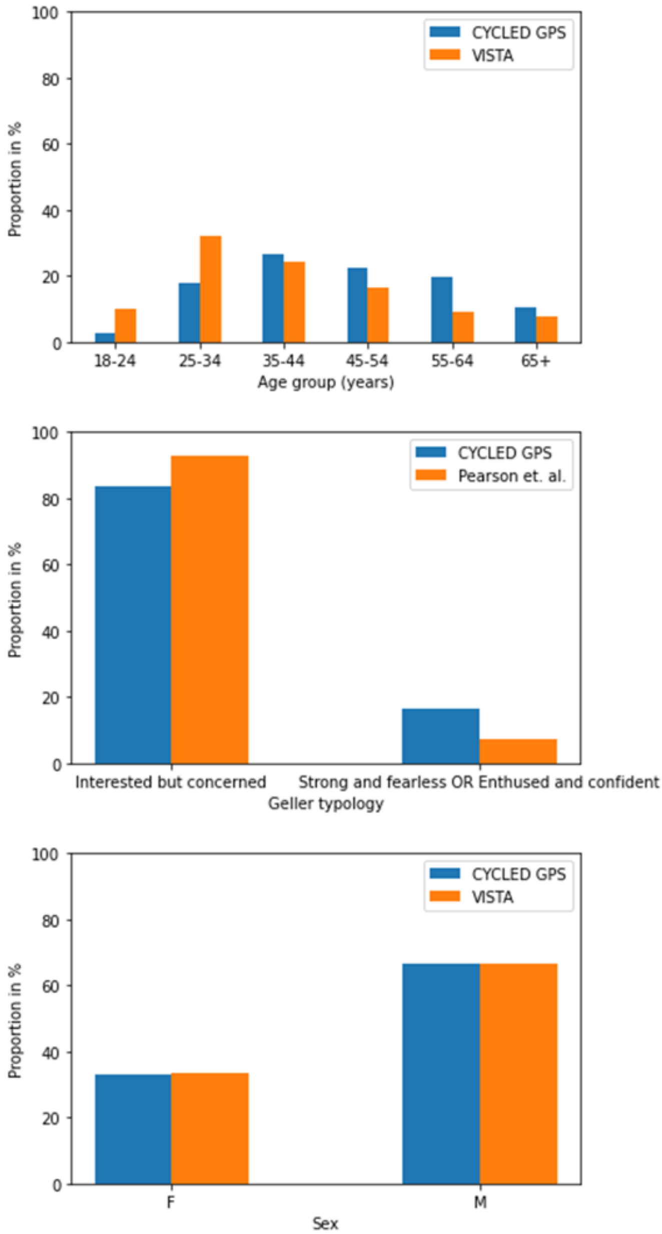



**Figure 2**: Distributions of demographic characteristics

that age group as well, with 35-44-year-old participants recording 5.6 bike trips per week on average. Adults less than 35 years (15.7%) old recorded

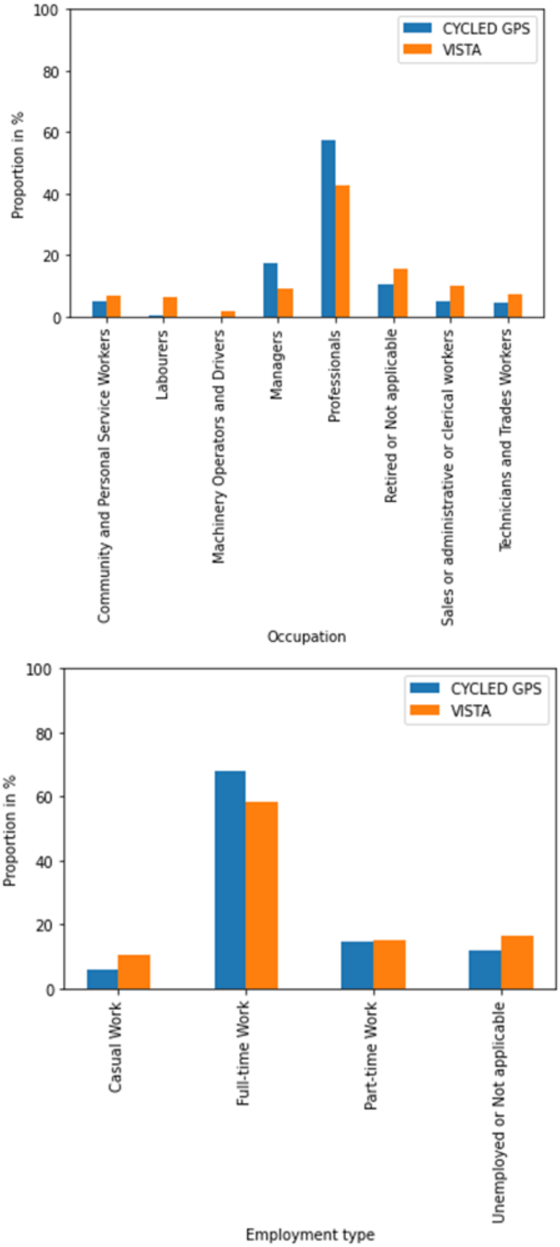
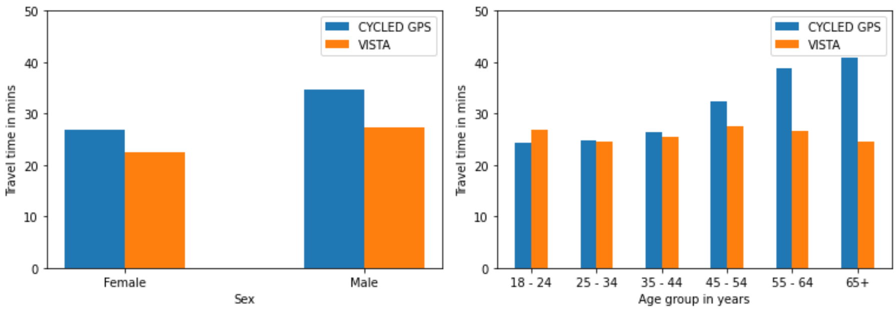


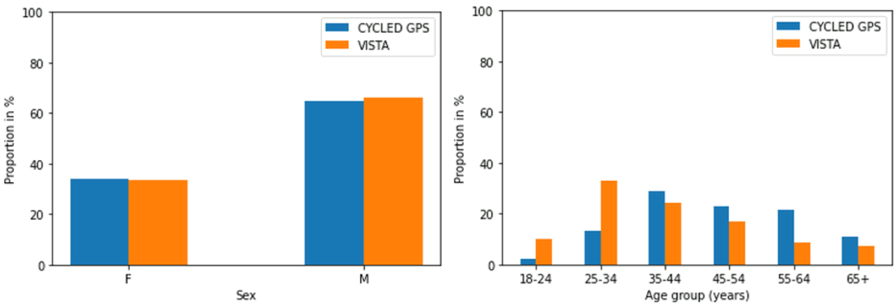

**Figure 3**: Distributions of demographic characteristics

fewer than 3 bike trips per week. As shown in Figure 5b, this distribution is slightly different from the underlying adult bike-riding population. Men not




(a) By sex (b) By age group

**Figure 4**: Travel time distribution by sex and age group




(a) By sex (b) By age group

**Figure 5**: Number of trips distribution by sex and age group

only recorded more trips than women but also more weekly trips than women (4.8 trips per week per person compared to 3.7 from women). Furthermore, bike trips made by men were significantly longer in terms of distance (10.6 km compared to 6.8 km, $U = 47924151$, $p \leq 0.001$) and duration (34.6 minutes compared to 26.9 minutes, $U = 45503333$, $p \leq 0.001$) than those made by women on average. Both the patterns conform to the underlying adult bike-riding population patterns as shown in Figure 4a and Figure 5a. Participants belonging to the 'Strong and Fearless' and 'Enthused and Confident' typologies put together recorded longer trips than those belonging to the 'Interested but Concerned' typology (10.6 km compared to 9.1 km, $U = 28623674$, $p \leq 0.001$).

**Table 2**: Demographic characteristics of participants

| Characteristic | Category | Participant Count (Percentage) | Population-level Percentages | $\chi^2$ [B] |
|---|---|---|---|---|
| Age | 18-24 years | 17 (2.5) | 10.2 | 187.9** |
| | 25-34 years | 120 (17.8) | 32.3 | |
| | 35-44 years | 180 (26.7) | 24.1 | |
| | 45-54 years | 152 (22.6) | 16.5 | |
| | 55-64 years | 133 (19.8) | 9.3 | |
| | 65+ years | 71 (10.6) | 7.6 | |
| Sex [C] | Female | 221 (32.8) | 33.3 | 0.006 |
| | Male | 445 (66.1) | 66.7 | |
| Geller typology [D] [E] | 'Strong and fearless' or 'Enthused and confident' | 109 (16.2) | 7.1 [F] | 83.5** |
| | 'Interested but concerned' | 562 (83.8) | 92.9 | |
| Occupation | Community and personal service | 32 (4.8) | 6.8 | 174.7** |
| | Labourers | 3 (0.4) | 6.4 | |
| | Machinery operators and drivers | 0 (0.0) | 1.9 | |
| | Manager | 116 (17.3) | 9.1 | |
| | Professional | 387 (57.6) | 42.6 | |
| | Retired or Not applicable | 71 (10.6) | 15.6 | |
| | Sales or administrative or clerical workers | 34 (5.1) | 10.1 | |
| | Technician and trades worker | 29 (4.3) | 7.4 | |
| Employment status | Full-time | 456 (67.8) | 58.3 | 32.3** |
| | Part-time | 99 (14.6) | 15.2 | |
| | Casual work | 39 (6.1) | 10.3 | |
| | Unemployment or Not applicable | 79 (11.7) | 6.1 | |
| Frequency of car usage | Daily | 119 (17.7) | | |
| | At least once a week but not daily | 380 (56.5) | | |
| | At least monthly but not weekly | 90 (13.4) | | |
| | Less than once per month | 48 (7.1) | | |
| | Never | 36 (5.3) | | |
| Frequency of bike usage | Daily | 181 (26.9) | | |
| | At least once a week but not daily | 462 (68.6) | | |
| | At least monthly but not weekly | 26 (3.9) | | |
| | Less than once per month | 3 (0.4) | | |
| Type of bike(s) owned | Pedal bike only | 596 (88.6) | | |
| | E-bike only | 67 (9.9) | | |
| | Both Pedal bike and E-bike | 10 (1.5) | | |

[B] *Statistically significance - *: $p \leq 0.05$, **: $p \leq 0.01$*

[C] *We had only 7 participants who neither identified as male nor female.*

[D] *We have merged 'Strong and fearless' with 'Enthused and confident' typologies.*

[E] *We have not presented 2 participants who were classified under the 'No way no how' typology but recorded bike trips.*

[F] *Obtained from a survey consisting of 3523 participants across Greater Melbourne [66]; Proportions recalculated after removing 'No way no how' cohort.*

**Table 3**: Preliminary trip statistics by age, gender and Geller typology

| **Characteristic** | **Category** | **Number of participants (Percentage)** | **Total number of bike trips detected (Percentage)** | **Number of bike trips detected per week per participant** | **Mean trip length (in kms)** | **Mean trip duration (in mins)** |
|---|---|---|---|---|---|---|
| **Age** | 18-24 years | 17 (2.5) | 446 (2.3) | 3.0 | 7.0 | 24.4 |
| | 25-34 years | 120 (17.8) | 2657 (13.4) | 2.8 | 7.0 | 24.8 |
| | 35-44 years | 180 (26.7) | 5698 (28.8) | 5.6 | 7.6 | 26.4 |
| | 45-54 years | 152 (22.6) | 4538 (22.9) | 4.8 | 10.1 | 32.5 |
| | 55-64 years | 133 (19.8) | 4247 (21.5) | 4.4 | 11.8 | 38.8 |
| | 65+ years | 71 (10.6) | 2196 (11.1) | 4.1 | 11.1 | 40.9 |
| **Gender** | Female | 221 (32.8) | 6699 (33.9) | 3.7 | 6.8 | 26.9 |
| | Male | 445 (66.1) | 12845 (64.9) | 4.8 | 10.6 | 34.6 |
| **Geller typology** | ‘Strong and fearless’ or ‘Enthused and confident’ | 109 (16.2) | 3344 (16.9) | 3.7 | 10.6 | 33.1 |
| | Interested but concerned | 562 (83.8) | 16423 (83.0) | 4.6 | 9.1 | 31.5 |

## 3.5 Spatial distribution

The spatial distribution of bicycling trips recorded in our data collection exercise is illustrated in Figure 6 at the network-level. As can be observed, there is a distinct pattern that most of our respondents collected data closer to the inner city and the Melbourne CBD such as the City of Melbourne (marked in the map), and the bicycling footprint reduces with distance from the CBD. This is reflective of population-level bicycling patterns in the Greater Melbourne region. [89]. We then made statistical comparisons between our sample trip

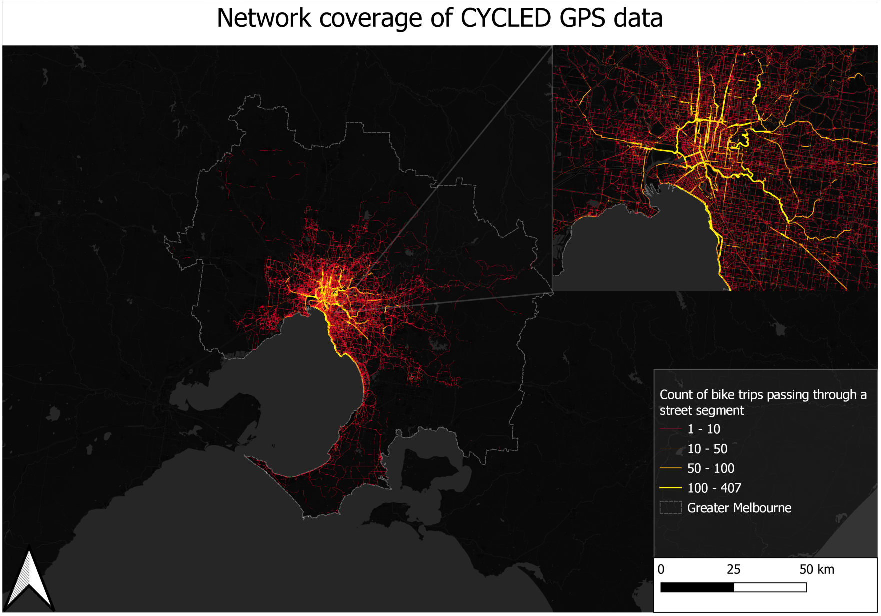



**Figure 6**: Coverage of the GPS data

origins and the population-level bike trip origins obtained from VISTA 2012-2020 data across the five clusters [67], as was mentioned in section 2.2.6. While the Chi-squared goodness of fit test showed that the distribution of our dataset was significantly different from the population-level dataset ($\chi^2 = 2231.9$, $p \leq 0.01$), the general patterns can be seen to be quite similar as illustrated in Figure 7 and Figure 8, except for the slight discrepancies between the outer west and the outer east. Statistically, we have compared the distribution of trips at the cluster-level, while graphically we have represented SA2-level map to facilitate better interpretation of the spatial representativeness. We did not conduct any statistical tests to ascertain spatial representativeness at an SA2-level.

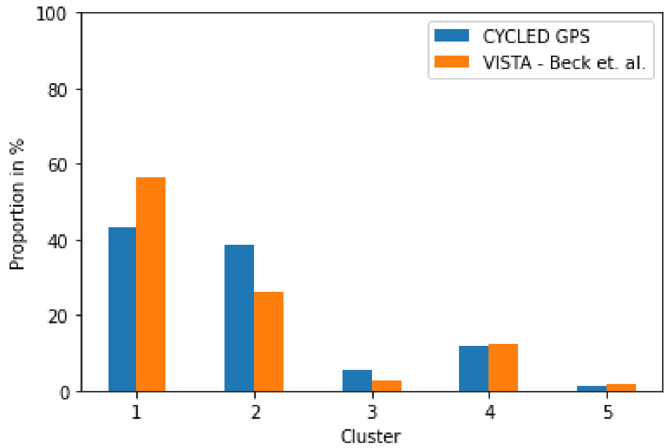



**Figure 7**: Distributions of bike trip origins across the five clusters

## 3.6 Infrastructure use

For an average bike trip, 14% and 10% of the trip length took place in arterial and collector road segments without any bike infrastructure respectively, while 35% of trip length took place on local roads with either painted bike lanes, sharrows or no bike infrastructure. 24% of an average bike trip was spent on offroad bike paths, while only 1% of an average bike trip took place on protected bike lanes. Details of infrastructure use is illustrated in Figure 9.

## 3.7 Temporal patterns

Figure 10 illustrates the distribution of starting times of bike trips of survey participants with corresponding bike-riding population-level estimates from the VISTA data. The distribution from our GPS dataset clearly replicates the population-level patterns with two distinct peaks, one in the morning (8-9 AM) and one in the evening (5-6 PM), with the fewest numbers observed towards the late night and early morning hours.

# 4 Discussion

In this section, we discuss the strengths of our data collection approach, the representativeness of our dataset, the utilities of our dataset, and limitations and future directions of the study.

## 4.1 Strengths of the data collection approach

While we collected bicycling GPS data using a smartphone application similar to most GPS data collection studies, we also made use of an associated Bluetooth beacon. Our method added the following values compared to other

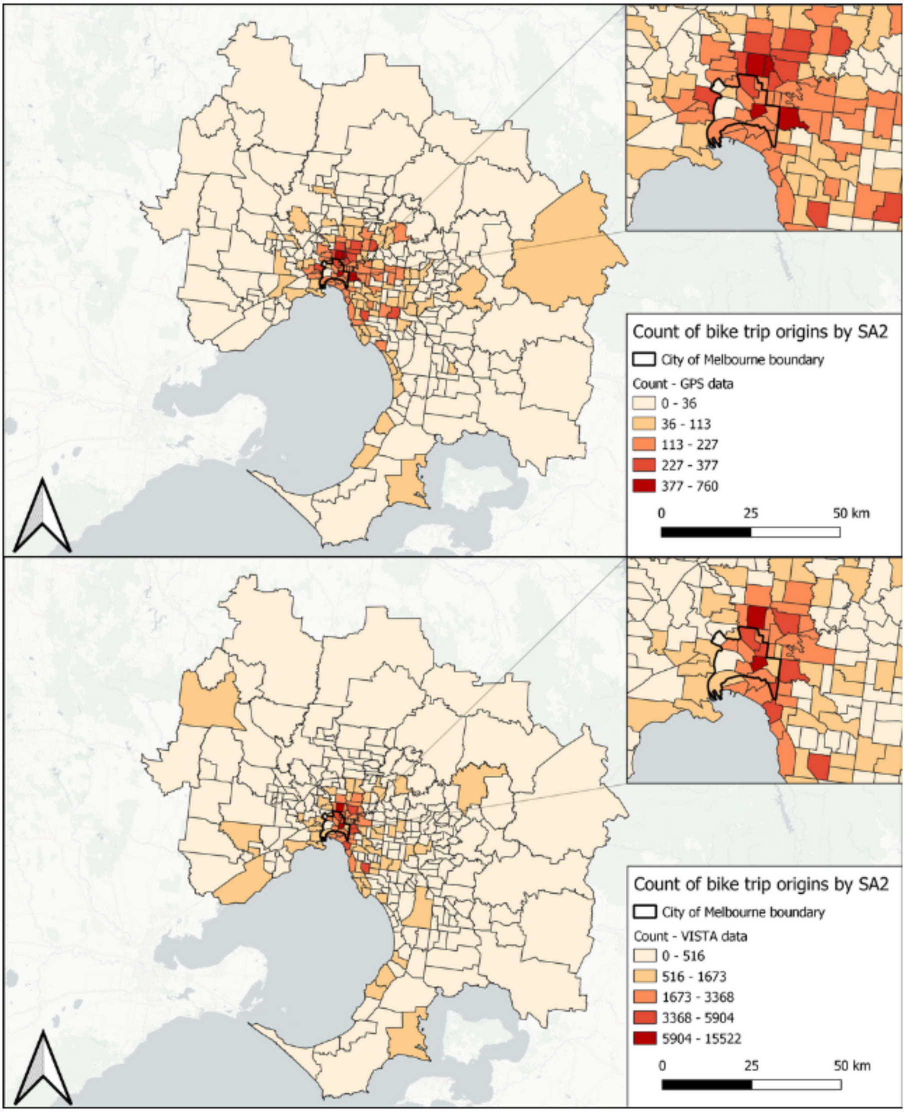



**Figure 8**: GPS data bike trip origins by SA2 (top) and VISTA 2012-2020 data bike trip origins by SA2 (bottom)

methods. First, our data collection method is optimised to reduce data preprocessing relative to existing methods. This was achieved by using a combination of a smartphone application and a Bluetooth beacon attached to a participant's bike, thereby limiting data collection to trips only taken by bike and not other travel modes, aided by minimal heuristic pre-processing. This is more efficient and accurate than previous methods that have either: 1) relied on people 'starting' and 'stopping' data collection (e.g. in a process similar to how a

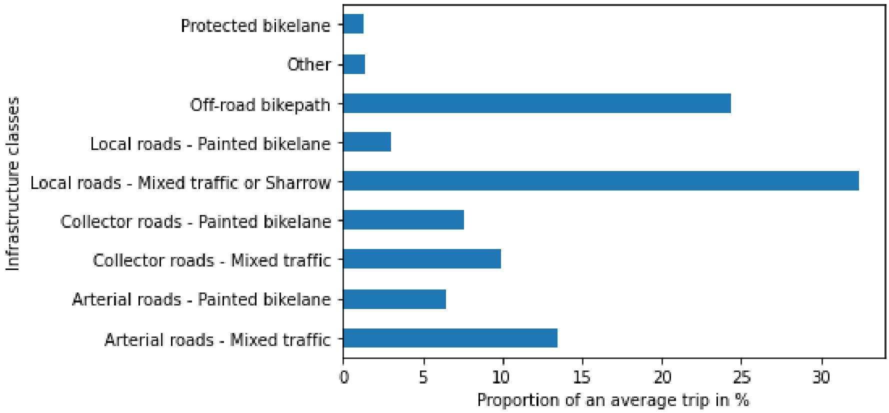



**Figure 9**: Infrastructure use by survey participants for an average bike trip

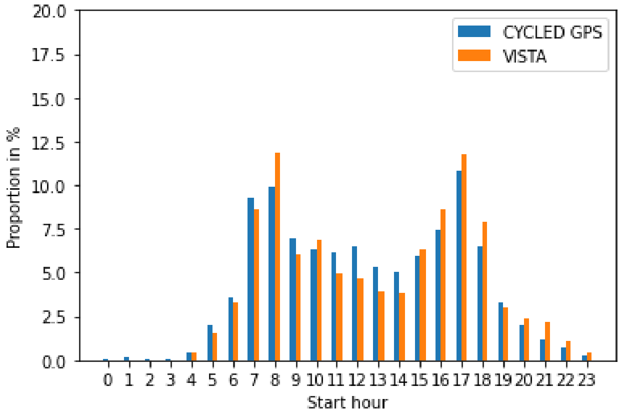



**Figure 10**: Distribution of starting times of bike trips

user may use the Strava application), which is a method subject to significant bias; or 2) relied on continuous GPS data collection without labelling of bicycle trips; this methods relies on mode-detection algorithms that suffer from inaccuracies. In our approach, we only needed to employ a simple and reliable heuristic-based approach for detecting bike trips. Second, this method allowed us to collect GPS data without requiring participant engagement to manually record bike trips. Thus, we were able to avoid self-reporting bias in our dataset by not missing out on bike trips that the participant could have forgotten to record, when the participants would remember midway through their bike trip to start recording GPS data, or when the participants would keep collecting data even after their trips were over. Third, GPS data collection is privacy-sensitive as it collects disaggregate-level location information and significantly consumes the smartphone battery used by the participant, both of which are major barriers to people partaking in similar studies or completing the defined duration of GPS data collection. Our data collection method ensured improved privacy as GPS data was only collected when people were on bicycles and not throughout the day. This potentially avoided significant participant dropout and proved critical for increased participation in our study. Therefore, our approach resulted in the collection of a large amount of bicycling GPS data (35.6 million GPS points, 19,782 bike trips from 673 users) across a large metropolitan area (9993 square kilometres) for a substantial period of eight months, spanning three seasons, which is not common for bicycling-specific GPS studies. Furthermore, our data processing methods have involved the use of OpenStreetMap, making it possible to be deployed across other locations around the globe.

## 4.2 Representativeness of the GPS data

Our innovative bicycling GPS data collection strategy coupled with our sampling approach resulted in the collection of a large dataset having sufficient coverage across multiple demographic subgroups and relevant bike-riding typologies. While we observed some statistically significant differences between the distributions between the distributions of our sample and the household travel survey, inspection of graphical plots demonstrated comparable distributions of demographic and trip characteristics. It must be noted that such comparisons with population-level distributions have never been reported in similar existing studies to the best of our knowledge. In the context of a field that has been plagued by the absence of robust and representative bicycle GPS datasets, the approach in our study of being able to quantify the representativeness of our sample relative to population representative samples is highly novel (and the first study to do so, to the best of our knowledge). We argue that our approach should become standard practice in the field to enhance the representativeness of sampling and transparency of reporting.

It was observed in Figure 6 that most of the bike trips were concentrated in the inner parts of Melbourne, taking place inside or near the City of Melbourne and that the number of trips declined as the distance from the inner-city area increased. A similar trend was observed while analysing the VISTA data which is the best representation of population-level bicycling behaviour in Greater Melbourne [89]. This trend is also similar to the spatial distribution of bike network density (total length of bikeable road network available divided by the area of the SA2). The majority of bike riders in Greater Melbourne (93%) belong to the typology 'Interested but Concerned' [66], reflecting people who feel comfortable and safe riding only in protected lanes or off-road paths [89]. The density of off-road bike paths and protected bike lanes across Greater Melbourne exhibits a pattern similar to the spatial variation of bike trips recorded in our dataset, as there is more bicycling infrastructure density in the inner city and this diminishes as the distance from the inner city increases. Corresponding illustrations are provided in Appendix D.

Younger adults were under-represented in our dataset, while non-adults (children) were not considered as part of this study as our GPS data collection focused on adult bike riders in Greater Melbourne. However, our methods are completely transferable to be applied to this important demographic group to understand their route choices and infrastructure needs. Therefore, our study provides a platform for further research related to 'bicycling to school' as it is an important consideration in city planning. In terms of Geller typologies, 83.5% of our participants belonged to the 'Interested but Concerned' group and recorded 83% of trips, which is interesting as riders from this group prefer to ride only in the presence of protected bike infrastructure (off-road paths and protected bike lanes), which makes up only 6.5% (2655 km) of bikeable street length across Greater Melbourne [89]. This is corroborated by the findings of a mixed methods study which stated that the assigned typologies did not always represent someone's confidence in riding a bike [90]. Given the high share of 'Interested but Concerned' participants, it will be interesting to understand whether their actual routes (as revealed by their GPS data) match their infrastructure preferences that define the Geller typology. This exploration was outside the scope of this study but will be investigated in subsequent studies.

## 4.3 Future research

### 4.3.1 Utility of the GPS data

#### *Bicyclist route choice modelling*

One of the objectives of this GPS data collection exercise is to develop bespoke route choice models (RCM) for Greater Melbourne [49, 91, 92]. Using the RCMs, we will have a deep understanding of the preferences regarding route characteristics of bike riders across Greater Melbourne, predict their behaviour across the transport network, and responses to changes in the network [10, 46, 93]. Given we were able to collect GPS data from a representative sample of adult bike riders, the generated RCM will produce results that are representative of the adult bike-riding population in Greater Melbourne which is key in the urban planning context. Furthermore, socio-demographic attributes of a rider, such as gender and age are significant drivers of route choice [94]. Therefore, we plan to not only develop a single RCM for Greater Melbourne, but multiple RCMs, one for each key population subgroup. This will help us understand how different subgroups of the bicycling population base their route choice decisions, and whether they are significantly different from each other (male vs female, younger vs older, experienced vs inexperienced rider). This will guide city councils in developing policies and introducing infrastructure that is more inclusive so that bicycling uptake can be significantly improved. In this regard, we acknowledge that our dataset contains both transport and leisure trips. Work is underway regarding classifying and filtering out leisure trips using algorithmic approaches, to develop route choice models with transport bike trips only.

***Modelling bicycling volumes***

Link-level bicycling volume estimates are essential for planners to understand bicycle flow dynamics at the finest spatial resolution [95] to strategically implement additional infrastructure or quantify the impact of infrastructure changes on individual roads within the network [27]. Link-level bicycling volume data is also necessary to appropriately measure cyclist safety (after accounting for exposure) on individual street segments. Existing bicycling volume models [95–98] have not always implemented robust, evidence-based bicycling route choice models. This is critical given that the route choices of bicyclists are vastly different from that of car drivers, with a greater focus on safety and separated bike infrastructure [39], and therefore needs careful consideration before its application to estimate link-level volumes. Future studies will develop more robust and representative bicycling volume models based on evidence-based RCMs using this GPS data.

***Other research questions***

Bicycling GPS data offers a host of other valuable objective insights into trends and patterns of bicycling activity which are useful information to support transport planning and policy-making. Infrastructure usage distribution is key to planners and policymakers and can only be reliably inferred from population-representative GPS datasets such as this. Future research will investigate the frequency of use of different types of bicycling infrastructure by our participants and investigate whether there are significant differences across population subgroups (men vs women). Furthermore, the data will be used to investigate measures such as operating speeds and travel time that are key indicators of perceived comfort and safety of bicyclists across different types of bicycling infrastructure, similar to studies conducted in Italy [99], Sweden [100], Korea [101], and the United States [102]. We will also evaluate the potential physical activity gains via GPS data by determining trips replaceable by bikes [103]. Also, with the growing popularity of electric bicycles or e-bikes, and with 77 e-bike riders among our survey participants, there lies an opportunity to investigate the differences between e-bike riders and non-e-bike riders in terms of their operating speeds, availed infrastructure, trip purpose, trip distance and detour tolerance, similar to studies conducted in the Netherlands [104, 105]. Our data will also help produce objective indicators for bicycling safety, which is one of the biggest barriers to bicycling uptake [4], such as measuring exposure and estimating crash risk on individual street segments, and across an entire urban area, similar to a Canadian study that mapped injury risk across Montreal by estimating Annual Average Daily Bicycling (AADB) volumes from GPS data [42].

### 4.3.2 Limitations and future directions

While this GPS dataset will deliver key insights into bicycling behaviour across Greater Melbourne, there exists certain limitations. While we started our recruitment strategy based on proportional stratified sampling, we switched to convenience sampling midway to tackle low participation rates, which was potentially underpinned by the logistical complexities associated with mailing and attaching a Bluetooth beacon. While our dataset represented the underlying population spatially and in terms of certain demographic characteristics, it was not able to significantly represent the underlying population-level Geller typology distribution. Opportunities exist in adopting approaches such as residual resampling and weighting in future that are used to address misrepresentation biases in mobility data [106, 107]. Nevertheless, we still collected significantly large amounts of data in terms of number of people and bike trips from a fairly representative sample.

While we are aware of alternative approaches to capture bike trips using smartphone GPS and Inertial Measurement Unit (IMU) data, such methods involve significant complexities as trip and mode detection algorithms need large amounts of labelled data, are often locally specific, not transferable [15], are dependent on the method of data collection, either do not focus or have lower accuracy for bicycling mode detection [75–78, 108], and are tested on small samples [109]. Therefore, to ensure the reliability of data collection and maximise the capture of bicycling trips without participant engagement and the aforementioned challenges, we chose to use Bluetooth beacons. It must be noted that despite using Bluetooth beacons to only capture bike rides of participants, we required a heuristic-based mode detection algorithm to remove a significant share of non-bike trips in pre-processing. This was discussed in detail in section 2.2.3. With our approach using the Bluetooth beacons, we were able to collect large-scale data. However, it must be noted that given the logistical challenges associated with distributing Bluetooth beacons to participants, our approach is logistically more challenging to scale up than GPS surveys not involving Bluetooth beacons.

Also, to minimise participant workload and considering the practical limitations of self-labelling trips from memory [12, 110], we did not ask participants to self-report their trip purposes. However, our data collection design could have included popup notifications asking participants to verify trip details after trip completion. Therefore, future study designs could consider integrating smartphone sensor-based data collection with a Bluetooth beacon, with the provision of immediate validation of trip details via self-reports that involve minimal participant intervention. The presence of a labelled dataset would have been beneficial for validating our trip detection and mode detection results. However, our algorithm heuristics and parameters were well-informed by existing literature and were calibrated to local conditions.

We acknowledge that our dataset contains both transport and leisure trips which have very distinct characteristics. However, we had strategically liaised with a diversity of organisations to support bicyclist recruitment to avoid the over-representation of recreational bike riders and to maximise the representativeness of our sample. At this moment, we have not differentiated between bicycling trips for transport and leisure as this was outside the scope of our immediate objectives. Furthermore, the comparisons were made with population-representative VISTA data which also included leisure bike-riding trips. Also, bicyclists are occasionally involved in multi-modal journeys in Greater Melbourne, where bicycles are used to access and being taken on public transport. While our current mode detection algorithm was not designed to identify such multi-modal bike trips, only 6.3% of bicyclists across Melbourne ride to access public transport (with even fewer carrying it on public transport) [111], and therefore does not undermine the results of this study. However, opportunities exist for future studies to account for multi-modal trips in future.

We also acknowledge that the comparisons between our GPS data collected in 2022 and VISTA 2012-2020 data are underpinned by a temporal mismatch. However, it is likely that 2012-2020 VISTA data (pre-COVID) is similar to 2022 cycling patterns as (a) current cycling participation rates are not significantly dissimilar to cycling rates during pre-COVID period (covered by VISTA) [111] and (b) 2012-2020 is the latest release of VISTA data. Finally, we used OpenStreetMap data for our data processing, the coverage and completeness of which are improving day by day, especially in developed countries such as Australia [112], and especially in urban areas such as Greater Melbourne [113]. However, it must be noted that OpenStreetMap is volunteered geographic information (VGI), and is, therefore, prone to occasional completeness and correctness issues [113], especially in the case of bicycling infrastructure due to inconsistent tagging practices [114].

# 5 Conclusion

Despite the usefulness of bicycling-specific GPS data for planning and policy-making purposes, large-scale data collection in urban areas has occurred significantly less frequently compared to its motorised counterparts, with data collection periods being shorter, spatial coverage being smaller, and studies seldom reporting the population-representativeness of their sample. We demonstrated the feasibility of collecting bicycling GPS data from a population-representative sample across a large spatial area using a novel bicycling GPS data collection system that allows for automatic capture of bike trips with minimal participant intervention. We collected GPS data from bike riders across Greater Melbourne, amassing a total of 19,782 trips from 673 participants across seven months with significant numbers from different population subgroups. Our data collection method involved pairing a smartphone application with a Bluetooth beacon attached to a participant's bike, thereby limiting data collection to trips only taken by bike, thus requiring minimal user interference, and mitigating self-reporting bias and extensive preprocessing. Our method reduced excessive data collection, thereby reducing privacy concerns among our participants, and reducing participant dropouts. We proposed an approach for capturing a population-representative sample and enabling quantification of representativeness, and our study sample was well-representative of the underlying bike-riding population, both spatially and demographically.

We presented details on the steps and methods that were adopted to process this data to prepare it for analysis, and extraction of meaningful information, and thus develop insights on trends and patterns of bicycling activity in Greater Melbourne. The collected dataset is shown to represent the underlying adult bike-riding population in Greater Melbourne fairly well, demographically and spatially. This population-representative dataset will be useful for planners and policymakers as it will assist in inferences on infrastructure usage and the development of models that will also be population-representative; something that is scarce in the bicycling GPS data collection domain. Such datasets have the potential to be used to develop robust route choice models to identify built-environment variables that significantly influence a rider's route choice, and consequently, to develop high-resolution bicycling volume estimates across large study areas, and advance our understanding of infrastructure utility, gender and typology differences, and the spatial distribution of bicyclists, thereby influencing evidence-based policy-making.

# References

[1] De Geus, B., De Smet, S., Nijs, J., Meeusen, R.: Determining the intensity and energy expenditure during commuter cycling. British Journal of Sports Medicine **41**(1), 8–12 (2007)

[2] Leyland, L.-A., Spencer, B., Beale, N., Jones, T., Van Reekum, C.M.: The effect of cycling on cognitive function and well-being in older adults. PloS one **14**(2), 0211779 (2019). https://doi.org/10.1371/journal.pone.0211779

[3] Lindsay, G., Macmillan, A., Woodward, A.: Moving urban trips from cars to bicycles: impact on health and emissions. Australian and New Zealand Journal of Public Health **35**(1), 54–60 (2011)

[4] Pearson, L., Gabbe, B., Reeder, S., Beck, B.: Barriers and enablers of bike riding for transport and recreational purposes in Australia. Journal of Transport & Health **28**, 101538 (2023). https://doi.org/10.1016/j.jth.2022.101538

[5] Pearson, L., Berkovic, D., Reeder, S., Gabbe, B., Beck, B.: Adults' self-reported barriers and enablers to riding a bike for transport: A systematic review. Transport Reviews **43**(3), 356–384 (2023). https://doi.org/10.1080/01441647.2022.2113570

[6] Roy, A., Nelson, T.A., Fotheringham, A.S., Winters, M.: Correcting bias in crowdsourced data to map bicycle ridership of all bicyclists. Urban Science **3**(2), 62 (2019). https://doi.org/10.3390/urbansci3020062

[7] Jestico, B., Nelson, T., Winters, M.: Mapping ridership using crowdsourced cycling data. Journal of Transport Geography **52**, 90–97 (2016). https://doi.org/10.1016/j.jtrangeo.2016.03.006

[8] Kwigizile, V., Oh, J.-S., Kwayu, K.: Integrating Crowdsourced Data with Traditionally Collected Data to Enhance Estimation of Bicycle Exposure Measure. Report, Western Michigan University (2019). https://wmich.edu/sites/default/files/attachments/u883/2019/TRCLC_RR_17_03.pdfhttps://wmich.edu/transportationcenter/trclc17-3https://trid.trb.org/view/1483416

[9] Huber, S., Lißner, S.: Disaggregation of aggregate GPS-based cycling data–How to enrich commercial cycling data sets for detailed cycling behaviour analysis. Transportation Research Interdisciplinary Perspectives **2**, 100041 (2019). https://doi.org/10.1016/j.trip.2019.100041

[10] Broach, J., Dill, J., Gliebe, J.: Where do cyclists ride? A route choice

model developed with revealed preference GPS data. Transportation Research Part A: Policy and Practice **46**(10), 1730–1740 (2012). https://doi.org/10.1016/j.tra.2012.07.005

[11] Charlton, B., Sall, E., Schwartz, M., Hood, J.: Bicycle route choice data collection using GPS-enabled smartphones. In: Transportation Research Board 90th Annual Meeting, 23-27 January 2011 (2011)

[12] Cich, G., Knapen, L., Bellemans, T., Janssens, D., Wets, G.: Trip/stop detection in gps traces to feed prompted recall survey. Procedia Computer Science **52**, 262–269 (2015). https://doi.org/10.1016/j.procs.2015.05.074. The 6th International Conference on Ambient Systems, Networks and Technologies (ANT-2015), the 5th International Conference on Sustainable Energy Information Technology (SEIT-2015)

[13] Keusch, F., Struminskaya, B., Antoun, C., Couper, M.P., Kreuter, F.: Willingness to participate in passive mobile data collection. Public Opinion Quarterly **83**(S1), 210–235 (2019). https://doi.org/10.1093/poq/nfz007

[14] Cottrill, C.D., Pereira, F.C., Zhao, F., Dias, I.F., Lim, H.B., Ben-Akiva, M.E., Zegras, P.C.: Future mobility survey: Experience in developing a smartphone-based travel survey in singapore. Transportation Research Record **2354**(1), 59–67 (2013)

[15] Lißner, S., Huber, S.: Facing the needs for clean bicycle data – a bicycle-specific approach of GPS data processing. European Transport Research Review **13**(1), 8 (2021). https://doi.org/10.1186/s12544-020-00462-2

[16] Kaya, S., Kilic, N., Kocak, T., Gungor, C.: A battery-friendly data acquisition model for vehicular speed estimation. Computers & Electrical Engineering **50**, 79–90 (2016). https://doi.org/10.1016/j.compeleceng.2016.01.017

[17] Lin, K., Kansal, A., Lymberopoulos, D., Zhao, F.: Energy-accuracy trade-off for continuous mobile device location. In: Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, pp. 285–298 (2010)

[18] Myr, D.: Traffic information gathering via cellular phone networks for intelligent transportation systems. Google Patents. US Patent 6,577,946 (2003)

[19] Lue, G., Miller, E.J.: Estimating a Toronto pedestrian route choice model using smartphone GPS data. Travel Behaviour and Society **14**, 34–42 (2019). https://doi.org/10.1016/j.tbs.2018.09.008

[20] Bolbol, A., Cheng, T., Tsapakis, I., Haworth, J.: Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. Computers, Environment and Urban Systems **36**(6), 526–537 (2012). https://doi.org/10.1016/j.compenvurbsys.2012.06.001

[21] Xiao, G., Juan, Z., Zhang, C.: Travel mode detection based on GPS track data and Bayesian networks. Computers, Environment and Urban Systems **54**, 14–22 (2015). https://doi.org/10.1016/j.compenvurbsys.2015.05.005

[22] Lee, K., Sener, I.N.: Emerging data for pedestrian and bicycle monitoring: Sources and applications. Transportation Research Interdisciplinary Perspectives **4**, 100095 (2020). https://doi.org/10.1016/j.trip.2020.100095

[23] Lißner, S., Huber, S., Lindemann, P., Anke, J., Francke, A.: GPS-data in bicycle planning:"Which cyclist leaves what kind of traces?" Results of a representative user study in Germany. Transportation Research Interdisciplinary Perspectives **7**, 100192 (2020). https://doi.org/10.1016/j.trip.2020.100192

[24] Hasan, R., Hasan, R.: Bluetooth low energy (BLE) beacon-based micro-positioning for pedestrians using smartphones in urban environments. Precision Positioning with Commercial Smartphones in Urban Environments, 135–149 (2021). https://doi.org/10.1007/978-3-030-71288-4_6

[25] Gunady, S., Keoh, S.L.: A non-GPS Based location tracking of public buses using Bluetooth proximity beacons. In: 2019 IEEE 5th World Forum on Internet of Things (WF-IoT), pp. 606–611 (2019). IEEE

[26] Ferreira, P., Zabolotny, A., Barreto, J.: Bicycle mode activity detection with bluetooth low energy beacons. In: 2019 IEEE 18th International Symposium on Network Computing and Applications (NCA), pp. 1–4 (2019). IEEE

[27] Bhowmick, D., Saberi, M., Stevenson, M., Thompson, J., Winters, M., Nelson, T., Leao, S.Z., Seneviratne, S., Pettit, C., Vu, H.L., et al.: A systematic scoping review of methods for estimating link-level bicycling volumes. Transport Reviews, 1–30 (2022). https://doi.org/10.1080/01441647.2022.2147240

[28] Lee, K., Sener, I.N.: Strava Metro data for bicycle monitoring: a literature review. Transport Reviews **41**(1), 27–47 (2021). https://doi.org/10.1080/01441647.2020.1798558

[29] Heesch, K.C., Langdon, M.: The usefulness of GPS bicycle tracking data for evaluating the impact of infrastructure change on cycling behaviour.

Health Promotion Journal of Australia **27**(3), 222–229 (2016). https://doi.org/10.1071/HE16032

[30] Lin, Z., Fan, W.: Bicycle ridership using crowdsourced data: Ordered probit model approach. Journal of Transportation Engineering Part A: Systems **146**(8), 04020076 (2020). https://doi.org/10.1061/JTEPBS.0000399

[31] Hong, J., McArthur, D.P., Stewart, J.L.: Can providing safe cycling infrastructure encourage people to cycle more when it rains? The use of crowdsourced cycling data (Strava). Transportation Research Part A: Policy and Practice **133**, 109–121 (2020). https://doi.org/10.1016/j.tra.2020.01.008

[32] Boss, D., Nelson, T., Winters, M., Ferster, C.J.: Using crowdsourced data to monitor change in spatial patterns of bicycle ridership. Journal of Transport & Health **9**, 226–233 (2018). https://doi.org/10.1016/j.jth.2018.02.008

[33] Dill, J., Gliebe, J.: Understanding and measuring bicycling behavior: A focus on travel time and route choice. Final report OTREC-RR-08-03 prepared for Oregon Transportation Research and Education Consortium (OTREC), Portland State University (2008)

[34] Goel, R., Goodman, A., Aldred, R., Nakamura, R., Tatah, L., Garcia, L.M.T., Zapata-Diomedi, B., de Sa, T.H., Tiwari, G., de Nazelle, A., Tainio, M., Buehler, R., Götschi, T., Woodcock, J.: Cycling behaviour in 17 countries across 6 continents: levels of cycling, who cycles, for what purpose, and how far? Transport Reviews **42**(1), 58–81 (2022). https://doi.org/10.1080/01441647.2021.1915898. doi: 10.1080/01441647.2021.1915898

[35] Pucher, J., Garrard, J., Greaves, S.: Cycling down under: a comparative analysis of bicycling trends and policies in Sydney and Melbourne. Journal of Transport Geography **19**(2), 332–345 (2011). https://doi.org/10.1016/j.jtrangeo.2010.02.007

[36] Vidal Tortosa, E., Lovelace, R., Heinen, E., Mann, R.P.: Cycling behaviour and socioeconomic disadvantage: An investigation based on the English National Travel Survey. Transportation Research Part A: Policy and Practice **152**, 173–185 (2021). https://doi.org/10.1016/j.tra.2021.08.004

[37] Buehler, R., Pucher, J.: International Overview: Cycling Trends in Western Europe, North America, and Australia. City Cycling, 9–29 (2012). https://doi.org/10.7551/mitpress/9434.003.0005

[38] Hoye, A.: Recommend or mandate? A systematic review and meta-analysis of the effects of mandatory bicycle helmet legislation. Accident Analysis & Prevention **120**, 239–249 (2018). https://doi.org/10.1016/j.aap.2018.08.001

[39] Winters, M., Teschke, K., Grant, M., Setton, E.M., Brauer, M.: How far out of the way will we travel? Built environment influences on route selection for bicycle and car travel. Transportation Research Record **2190**(1), 1–10 (2010)

[40] Reddy, S., Shilton, K., Denisov, G., Cenizal, C., Estrin, D., Srivastava, M.: Biketastic: sensing and mapping for better biking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1817–1820 (2010). https://doi.org/10.1145/1753326.1753598

[41] Broach, J., Gliebe, J., Dill, J.: Bicycle route choice model developed using revealed preference GPS data. In: 90th Annual Meeting of the Transportation Research Board, Washington, DC (2011)

[42] Strauss, J., Miranda-Moreno, L.F., Morency, P.: Mapping cyclist activity and injury risk in a network combining smartphone GPS data and bicycle counts. Accident Analysis and Prevention **83**, 132–142 (2015). https://doi.org/10.1016/j.aap.2015.07.014

[43] Romanillos, G., Zaltz Austwick, M.: Madrid cycle track: Visualizing the cyclable city. Journal of Maps **12**(5), 1218–1226 (2016). https://doi.org/10.1080/17445647.2015.1088901

[44] Harvey, F.J., Krizek, K.J.: Commuter bicyclist behavior and facility disruption. Technical report, University of Minnesota (2007)

[45] Hood, J., Sall, E., Charlton, B.: A GPS-based bicycle route choice model for San Francisco, California. Transportation Letters: The International Journal of Transportation Research **3**(1), 63–75 (2011)

[46] Chen, P., Shen, Q., Childress, S.: A GPS data-based analysis of built environment influences on bicyclist route preferences. International Journal of Sustainable Transportation **12**(3), 218–231 (2018). https://doi.org/10.1080/15568318.2017.1349222

[47] Hudson, J.G., Duthie, J.C., Rathod, Y.K., Larsen, K.A., Meyer, J.L., et al.: Using smartphones to collect bicycle travel data in Texas. Technical report, Texas Transportation Institute. University Transportation Center for Mobility (2012)

[48] Park, Y., Akar, G.: Why do bicyclists take detours? A multilevel regression model using smartphone GPS data. Journal of Transport Geography

**74**, 191–200 (2019). https://doi.org/10.1016/j.jtrangeo.2018.11.013

[49] Menghini, G., Carrasco, N., Schüssler, N., Axhausen, K.W.: Route choice of cyclists in Zurich. Transportation Research Part A: Policy and Practice **44**(9), 754–765 (2010). https://doi.org/10.1016/j.tra.2010.07.008

[50] Łukawska, M., Paulsen, M., Rasmussen, T.K., Jensen, A.F., Nielsen, O.A.: A joint bicycle route choice model for various cycling frequencies and trip distances based on a large crowdsourced GPS dataset. Transportation Research Part A: Policy and Practice **176**, 103834 (2023). https://doi.org/10.1016/j.tra.2023.103834

[51] van de Coevering, P., de Kruijf, J., Bussche, D.: Bike print. Policy renewal and innovation by means of tracking technology. In: Colloquium Vervoersplanologisch Speurwerk, Eindhoven, Netherlands (2014)

[52] Oskarbski, J., Birr, K., Żarski, K.: Bicycle traffic model for sustainable urban mobility planning. Energies **14**(18), 5970 (2021). https://doi.org/10.3390/en14185970

[53] Rupi, F., Poliziani, C., Schweizer, J.: Data-driven Bicycle Network Analysis Based on Traditional Counting Methods and GPS Traces from Smartphone. ISPRS International Journal of Geo-Information **8**(8), 322 (2019). https://doi.org/10.3390/ijgi8080322

[54] Poliziani, C., Rupi, F., Mbuga, F., Schweizer, J., Tortora, C.: Categorizing three active cyclist typologies by exploring patterns on a multitude of GPS crowdsourced data attributes. Research in Transportation Business & Management **40**, 100572 (2021). https://doi.org/10.1016/j.rtbm.2020.100572

[55] Pritchard, R., Bucher, D., Frøyen, Y.: Does new bicycle infrastructure result in new or rerouted bicyclists? A longitudinal GPS study in Oslo. Journal of Transport Geography **77**, 113–125 (2019). https://doi.org/10.1016/j.jtrangeo.2019.05.005

[56] Ton, D., Duives, D., Cats, O., Hoogendoorn, S.: Evaluating a data-driven approach for choice set identification using GPS bicycle route choice data from Amsterdam. Travel Behaviour and Society **13**, 105–117 (2018). https://doi.org/10.1016/j.tbs.2018.07.001

[57] Pritchard, R.: Revealed preference methods for studying bicycle route choice—A systematic review. International journal of Environmental Research and Public Health **15**(3), 470 (2018)

[58] Łukawska, M.: Quantitative modelling of cyclists' route choice behaviour

on utilitarian trips based on gps data: associated factors and behavioural implications. Transport Reviews, 1–32 (2024). https://doi.org/10.1080/01441647.2024.2355468. doi: 10.1080/01441647.2024.2355468

[59] Molloy, J., Castro, A., Götschi, T., Schoeman, B., Tchervenkov, C., Tomic, U., Hintermann, B., Axhausen, K.W.: The MOBIS dataset: a large GPS dataset of mobility behaviour in Switzerland. Transportation **50**(5), 1983–2007 (2023). https://doi.org/10.1007/s11116-022-10299-4

[60] Pettit, C.J., Lieske, S.N., Leao, S.Z.: Big bicycle data processing: From personal data to urban applications. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences **3**, 173 (2016)

[61] Leao, S.Z., Pettit, C.: Mapping bicycling patterns with an agent-based model, census and crowdsourced data. In: Agent Based Modelling of Urban Systems: First International Workshop, ABMUS 2016, Singapore, pp. 112–128 (2017). Springer

[62] Krizek, K.J., Johnson, P.J., Tilahun, N., *et al.*: Gender differences in bicycling behavior and facility preferences. Research on Women's Issues in Transportation **2**, 31–40 (2005)

[63] Sustainable Mobility and Safety Research Group, Monash University: Bicycling infrastructure classification using OpenStreetMap (2023). https://doi.org/10.5281/zenodo.8274978. https://github.com/SustainableMobility/bicycling-infrastructure-classification

[64] Dorofeev, S., Grant, P.: Statistics for Real-life Sample Surveys: Non-simple-random Samples and Weighted Data. Cambridge University Press, Cambridge (2006). https://doi.org/10.1017/CBO9780511543265

[65] Department of Transport and Planning: Victorian Integrated Survey of Travel and Activity. https://dtp.vic.gov.au/about/data-and-research/vista (2022)

[66] Pearson, L., Dipnall, J., Gabbe, B., Braaf, S., White, S., Backhouse, M., Beck, B.: The potential for bike riding across entire cities: Quantifying spatial variation in interest in bike riding. Journal of Transport & Health **24**, 101290 (2022). https://doi.org/10.1016/j.jth.2021.101290

[67] Beck, B., Winters, M., Nelson, T., Pettit, C., Leao, S.Z., Saberi, M., Thompson, J., Seneviratne, S., Nice, K., Stevenson, M.: Developing urban biking typologies: Quantifying the complex interactions of bicycle ridership, bicycle network and built environment characteristics. Environment and Planning B: Urban Analytics and City Science **50**(1), 7–23 (2023). https://doi.org/10.1177/23998083221100827

[68] Geller, R.: Four Types of Cyclists. Portland Bureau of Transportation (2006)

[69] Zheng, Y.: Trajectory data mining: an overview. ACM Transactions on Intelligent Systems and Technology (TIST) **6**(3), 1–41 (2015)

[70] Lin, K., Xu, Z., Qiu, M., Wang, X., Han, T.: Noise filtering, trajectory compression and trajectory segmentation on GPS data. In: 2016 11th International Conference on Computer Science & Education (ICCSE), pp. 490–495 (2016). https://doi.org/10.1109/ICCSE.2016.7581629

[71] Naumov, V., Banet, K.: Estimating parameters of demand for trips by public bicycle system using GPS data. In: Smart and Green Solutions for Transport Systems: 16th Scientific and Technical Conference" Transport Systems. Theory and Practice 2019" Selected Papers 16, pp. 213–224 (2020). Springer

[72] Wolf, J., Schönfelder, S., Samaga, U., Oliveira, M., Axhausen, K.W.: Eighty weeks of global positioning system traces: approaches to enriching trip information. Transportation Research Record **1870**(1), 46–54 (2004)

[73] Bhowmick, D., Winter, S., Stevenson, M., Vortisch, P.: The impact of urban road network morphology on pedestrian wayfinding behavior. Journal of Spatial Information Science (21), 203–228 (2020). https://doi.org/10.5311/JOSIS.2020.21.601

[74] Schuessler, N., Axhausen, K.W.: Processing raw data from global positioning systems without additional information. Transportation Research Record **2105**(1), 28–36 (2009)

[75] Gong, H., Chen, C., Bialostozky, E., Lawson, C.T.: A GPS/GIS method for travel mode detection in New York City. Computers, Environment and Urban Systems **36**(2), 131–139 (2012). https://doi.org/10.1016/j.compenvurbsys.2011.05.003

[76] Prelipcean, A.C., Gidófalvi, G., Susilo, Y.O.: Transportation mode detection–an in-depth review of applicability and reliability. Transport Reviews **37**(4), 442–464 (2017). https://doi.org/10.1080/01441647.2016.1246489

[77] Shin, D.: Urban sensing by crowdsourcing: Analysing urban trip behaviour in Zurich. International Journal of Urban and Regional Research **40**(5), 1044–1060 (2016)

[78] Zhang, L., Dalyot, S., Eggert, D., Sester, M.: Multi-stage approach to travel-mode segmentation and classification of GPS traces. International

Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences: Geospatial Data Infrastructure: From Data Acquisition And Updating To Smarter Services **38**(W25), 87–93 (2011)

[79] Meert, W., Verbeke, M.: HMM with non-emitting states for Map Matching. In: European Conference on Data Analysis (ECDA), Paderborn, Germany (2018)

[80] Newson, P., Krumm, J.: Hidden Markov map matching through noise and sparseness. In: Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, pp. 336–343 (2009)

[81] Saki, S., Hagen, T.: A Practical Guide to an Open-Source Map-Matching Approach for Big GPS Data. SN Computer Science **3**(5), 415 (2022). https://doi.org/10.1007/s42979-022-01340-5

[82] Wu, H., Huang, S., Fu, C., Xu, S., Wang, J., Huang, W., Liu, C.: Online map-matching assisted by object-based classification of driving scenario. International Journal of Geographical Information Science **37**(8), 1872–1907 (2023). https://doi.org/10.1080/13658816.2023.2206877. doi: 10.1080/13658816.2023.2206877

[83] Trogh, J., Botteldooren, D., Coensel, B.D., Martens, L., Joseph, W., Plets, D.: Map matching and lane detection based on markovian behavior, gis, and imu data. IEEE Transactions on Intelligent Transportation Systems **23**(3), 2056–2070 (2022). https://doi.org/10.1109/TITS.2020.3031080

[84] Javanmardi, E., Javanmardi, M., Gu, Y., Kamijo, S.: Pre-estimating self-localization error of ndt-based map-matching from map only. IEEE Transactions on Intelligent Transportation Systems **22**(12), 7652–7666 (2021). https://doi.org/10.1109/TITS.2020.3006854

[85] OpenStreetMap contributors: OpenStreetMap. https://www.openstreetmap.org (2022)

[86] Boeing, G.: OSMnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Computers, Environment and Urban Systems **65**, 126–139 (2017). https://doi.org/10.1016/j.compenvurbsys.2017.05.00

[87] Australian Bureau of Statistics: Statistical Area Level 2. ABS. https://www.abs.gov.au/statistics/standards/australian-statistical-geography-standard-asgs-edition-3/jul2021-jun2026/main-structure-and-greater-capital-city-statistical-areas/statistical-area-level-2.

[88] Sustainable Mobility and Safety Research Group, Monash University: Highway/Road classification using OpenStreetMap (2024). https://doi.org/10.6084/m9.figshare.27059980.v1. https://github.com/SustainableMobility/highway-classification-osm

[89] Beck, B., Winters, M., Thompson, J., Stevenson, M., Pettit, C.: Spatial variation in bicycling: A retrospective review of travel survey data from Greater Melbourne, Australia. SocArXiv (2021). https://doi.org/10.31235/osf.io/78qgf

[90] Hosford, K., Laberee, K., Fuller, D., Kestens, Y., Winters, M.: Are they really interested but concerned? A mixed methods exploration of the Geller bicyclist typology. Transportation Research Part F: Traffic Psychology and Behaviour **75**, 26–36 (2020)

[91] Zimmermann, M., Mai, T., Frejinger, E.: Bike route choice modeling using GPS data without choice sets of paths. Transportation Research Part C: Emerging Technologies **75**, 183–196 (2017). https://doi.org/10.1016/j.trc.2016.12.009

[92] Yeboah, G., Alvanides, S.: Route choice analysis of urban cycling behaviors using openstreetmap: Evidence from a british urban environment. OpenStreetMap in GIScience: Experiences, Research, and Applications, 189–210 (2015). https://doi.org/10.1007/978-3-319-14280-7_10

[93] Prato, C.G.: Route choice modeling: past, present and future research directions. Journal of Choice Modelling **2**(1), 65–100 (2009)

[94] Rupi, F., Freo, M., Poliziani, C., Postorino, M.N., Schweizer, J.: Analysis of gender-specific bicycle route choices using revealed preference surveys based on GPS traces. Transport Policy (2023). https://doi.org/10.1016/j.tranpol.2023.01.001

[95] Kaziyeva, D., Loidl, M., Wallentin, G.: Simulating spatio-temporal patterns of bicycle flows with an agent-based model. ISPRS International Journal of Geo-Information **10**(2) (2021). https://doi.org/10.3390/ijgi10020088

[96] Wallentin, G., Loidl, M.: Agent-based bicycle traffic model for salzburg city. GI Forum – Journal for Geographic Information Science **3**, 558–566 (2015). https://doi.org/10.1553/giscience2015s558

[97] Jacyna, M., Wasiak, M., Kłodawski, M., Golebiowski, P.: Modelling of bicycle traffic in the cities using visum. In: 10th International Scientific Conference Transbaltica 2017: Transportation Science and Technology, vol. 187, pp. 435–441. https://doi.org/10.1016/j.proeng.2017.04.397

[98] Gosse, C.A., Clarens, A.: Estimating spatially and temporally continuous bicycle volumes by using sparse data. Transportation Research Record **2443**, 115–122 (2014). https://doi.org/10.3141/2443-13

[99] Poliziani, C., Rupi, F., Schweizer, J.: Traffic surveys and GPS traces to explore patterns in cyclist's in-motion speeds. Transportation Research Procedia **60**, 410–417 (2022). https://doi.org/10.1016/j.trpro.2021.12.053

[100] Manum, B., Arnesen, P., Nordstrom, T., Gil, J.: Improving GIS-based models for bicycling speed estimations. Transportation Research Procedia **42**, 85–99 (2019). https://doi.org/10.1016/j.trpro.2019.12.009

[101] Joo, S., Oh, C., Jeong, E., Lee, G.: Categorizing bicycling environments using GPS-based public bicycle speed data. Transportation Research Part C: Emerging Technologies **56**, 239–250 (2015). https://doi.org/10.1016/j.trc.2015.04.012

[102] El-geneidy, A., Krizek, K.J., Iacono, M.J.: Predicting Bicycle Travel Speeds Along Different Facilities Using GPS Data: A Proof-of-Concept Model. In: Transportation Research Board 86th Annual Meeting, 21-25 January 2007 (2007)

[103] Loh, V., Sahlqvist, S., Veitch, J., Thornton, L., Salmon, J., Cerin, E., Schipperijn, J., Timperio, A.: From motorised to active travel: using GPS data to explore potential physical activity gains among adolescents. BMC Public Health **22**(1), 1–9 (2022). https://doi.org/10.1186/s12889-022-13947-7

[104] Plazier, P.A., Weitkamp, G., van den Berg, A.E.: "Cycling was never so easy!" An analysis of e-bike commuters' motives, travel behaviour and experiences using GPS-tracking and interviews. Journal of Transport Geography **65**, 25–34 (2017). https://doi.org/10.1016/j.jtrangeo.2017.09.017

[105] Dane, G., Feng, T., Luub, F., Arentze, T.: Route choice decisions of E-bike users: Analysis of GPS tracking data in the Netherlands. In: Geospatial Technologies for Local and Regional Development: Proceedings of the 22nd AGILE Conference on Geographic Information Science 22, pp. 109–124 (2020). https://doi.org/10.1007/978-3-030-14745-7_7. Springer

[106] Pappalardo, L., Manley, E., Sekara, V., Alessandretti, L.: Future directions in human mobility science. Nature Computational Science **3**(7), 588–600 (2023). https://doi.org/10.1038/s43588-023-00469-4

[107] Schlosser, F., Sekara, V., Brockmann, D., Garcia-Herranz, M.: Biases in

human mobility data impact epidemic modeling (2021)

[108] Berjisian, E., Bigazzi, A.: Evaluation of methods to distinguish trips from activities in walking and cycling GPS data. Transportation Research Part C: Emerging Technologies **137**, 103588 (2022). https://doi.org/10.1016/j.trc.2022.103588

[109] Nikolic, M., Bierlaire, M.: Review of transportation mode detection approaches based on smartphone data. In: 17th Swiss Transport Research Conference, Monte Verità / Ascona (2017)

[110] Fillekes, M.P., Kim, E.-K., Trumpf, R., Zijlstra, W., Giannouli, E., Weibel, R.: Assessing Older Adults' Daily Mobility: A Comparison of GPS-Derived and Self-Reported Mobility Indicators. Sensors **19**(20), 4551 (2019). https://doi.org/10.3390/s19204551

[111] Bolton, S.: National Walking and Cycling Participation Survey 2023. Technical report, Cycling and Walking Australia and New Zealand (CWANZ) (2023)

[112] Arsanjani, J., Zipf, A., Mooney, P., Helbich, M.: An introduction to OpenStreetMap in Geographic Information Science: Experiences, research, and applications. OpenStreetMap in GIScience, 1–15 (2015)

[113] Ferster, C., Fischer, J., Manaugh, K., Nelson, T., Winters, M.: Using openstreetmap to inventory bicycle infrastructure: A comparison with open data from cities. International Journal of Sustainable Transportation **14**(1), 64–73 (2020). https://doi.org/10.1080/15568318.2018.1519746

[114] Vierø, A.R., Vybornova, A., Szell, M.: Bikedna: A tool for bicycle infrastructure data and network assessment. Environment and Planning B: Urban Analytics and City Science, 23998083231184471 (2023). https://doi.org/10.1177/23998083231184471

# Appendix A Trip detection algorithm

This step involved identifying separate trips from the entire GPS dataset of an individual. Therefore, trip detection is also referred to as trajectory segmentation. We developed our algorithm in this regard, based on our unique data collection method [23, 69, 71]. The algorithm involved detection of temporal gaps and stay locations in the data [69, 74], explained as follows. First, we calculated the time difference between consecutive points in an individual's dataset. We labelled these time differences as a temporal gap when it was more than or equal to 600 seconds. Second, we applied the stop detection algorithm developed by *scikit-mobility* which uses spatial clustering techniques to identify stops or stay locations within a given segment. We set the spatial radius at 100 metres [69] and the temporal threshold as 600 seconds, meaning that if the user did not move beyond a 100 metres of the first point of a dynamic 600-second window, the user is considered to be stationary for all this time. After passing 600 seconds, the stop is considered to have ended when the user's location is detected beyond 100 metres of the first point. We recorded the start and end times of the stay locations that were detected. We combined the start and end time information from temporal gaps and stay locations for a user, and based on this, we segmented the entire trajectory data into meaningful segments. Third, we removed segments which were less than 60 seconds or comprised of less than 30 data points. VISTA data reports trips which have a duration of at least one minute. Also, it is extremely less likely for bicycling trips less than one minute to contain any meaningful or representative information. For the same reason, we removed trips which comprised of less than 30 data points. The fourth and final step involved applying the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm to detect and remove stationary segments (that were not detected as stay locations by the *scikit-mobility's* stay location detection algorithm). Here clustered points represented the user being stationary while the noise points represented the movement of the user. We removed segments which had only one cluster or when the number of noise points were less than 25 or less than 5% of the clustered points in a segment. At the end, we conducted a Point-in-Polygon analysis to check if the origin (first GPS point of the trip) or the destination (the final GPS point of the trip) fell within the boundaries of Greater Melbourne, otherwise we discarded the trip. Finally, we considered the segments that remained in our dataset as trips for further analysis.

# Appendix B Mode detection flowchart

# Appendix C Additional data tables

**Table C1**: Preliminary trip statistics by age and gender

| Age | Gender | Number of participants | Total number of bike trips detected | Number of bike trips detected per week per participant |
|---|---|---|---|---|
| 18-24 years | Female | 4 | 86 | 2.02 |
| 25-34 years | | 50 | 1155 | 2.09 |
| 35-44 years | | 71 | 2498 | 3.86 |
| 45-54 years | | 47 | 1591 | 3.34 |
| 55-64 years | | 40 | 1342 | 3.78 |
| 65+ years | | 19 | 580 | 3.28 |
| 18-24 years | Male | 14 | 377 | 3.25 |
| 25-34 years | | 73 | 1774 | 2.57 |
| 35-44 years | | 116 | 3558 | 3.87 |
| 45-54 years | | 107 | 3292 | 3.44 |
| 55-64 years | | 98 | 3271 | 3.33 |
| 65+ years | | 54 | 1858 | 3.66 |

**Table C2**: Preliminary trip statistics by age and Geller typology

| Age | Geller typology | Number of participants | Total number of bike trips detected | Number of bike trips detected per week per participant |
|---|---|---|---|---|
| 18-24 years | 'Strong and fearless' or 'Enthused and confident' | 3 | 134 | 4.18 |
| 25-34 years | | 22 | 523 | 2.38 |
| 35-44 years | | 27 | 1022 | 4.5 |
| 45-54 years | | 34 | 1228 | 3.45 |
| 55-64 years | | 16 | 423 | 2.48 |
| 65+ years | | 8 | 312 | 3.9 |
| 18-24 years | Interested but concerned | 15 | 331 | 2.74 |
| 25-34 years | | 102 | 2422 | 2.37 |
| 35-44 years | | 160 | 5149 | 3.84 |
| 45-54 years | | 122 | 3750 | 3.42 |
| 55-64 years | | 123 | 4201 | 3.57 |
| 65+ years | | 65 | 2126 | 3.52 |

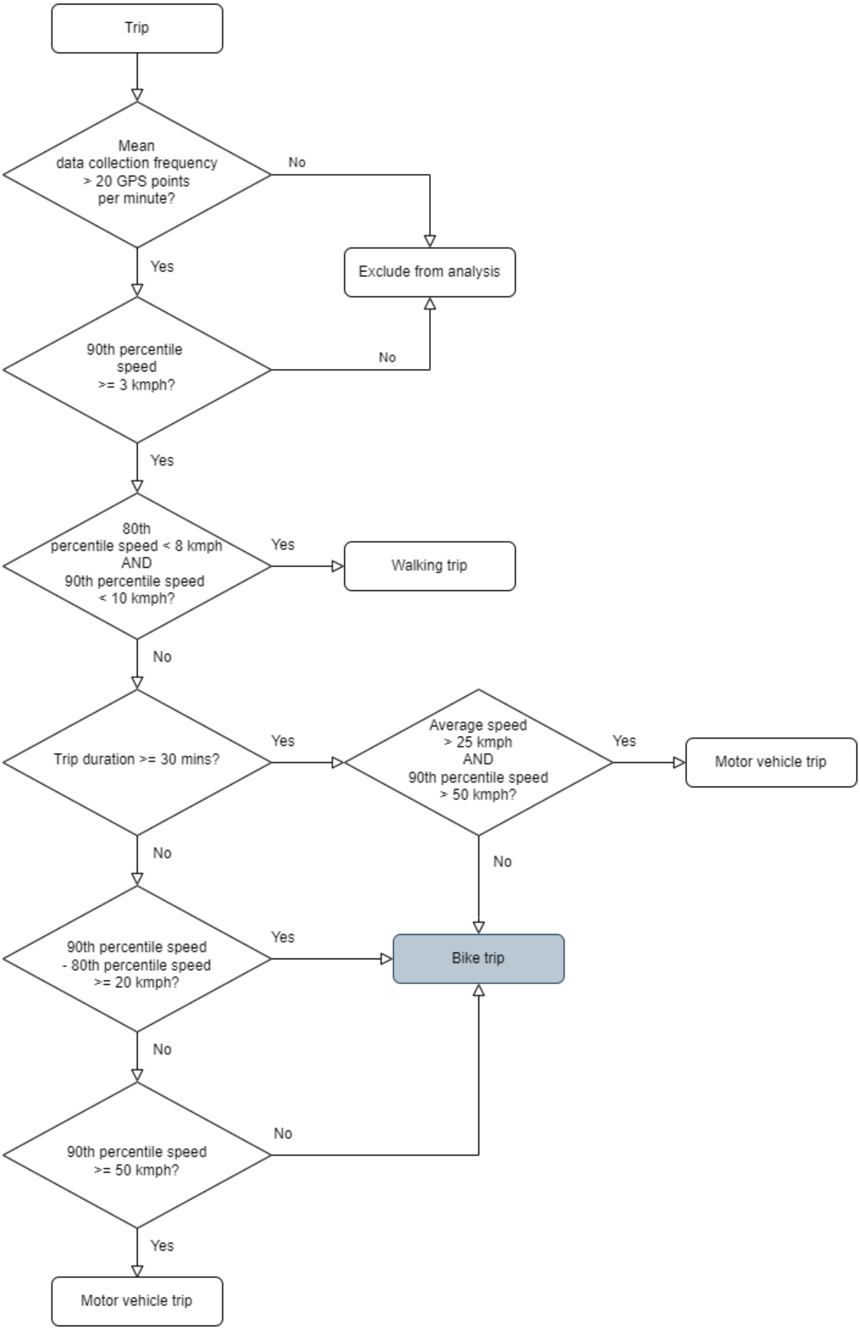



**Figure B1**: Mode detection flowchart

**Table C3**: Trip speed statistics by age and gender

| Age | Gender | Number of participants | Mean trip speed (in kmph) | Mean 20th percentile speed (in kmph) | Mean 80th percentile speed (in kmph) | Mean 90th percentile speed (in kmph) |
|---|---|---|---|---|---|---|
| 18-24 | Female | 4 | 13.3 | 6.15 | 25.41 | 28.59 |
| 25-34 | | 48 | 14.3 | 3.62 | 21.66 | 25.99 |
| 35-44 | | 68 | 14.8 | 5.03 | 22.44 | 25.39 |
| 45-54 | | 44 | 15.1 | 6.82 | 24.21 | 27.87 |
| 55-64 | | 38 | 16.0 | 7.47 | 24.9 | 28.19 |
| 65+ | | 19 | 14.8 | 6.31 | 20.78 | 24.24 |
| 18-24 | Male | 12 | 18.4 | 5.65 | 22.58 | 26.4 |
| 25-34 | | 70 | 18.3 | 7.09 | 25.53 | 29.55 |
| 35-44 | | 111 | 16.9 | 7.39 | 25.23 | 29.11 |
| 45-54 | | 106 | 18.5 | 7.95 | 26.86 | 31.19 |
| 55-64 | | 94 | 17.3 | 8.09 | 25.06 | 28.73 |
| 65+ | | 52 | 16.2 | 7.38 | 23.92 | 27.51 |

**Table C4**: Trip speed statistics by Geller typology

| Geller typology | Number of participants | Mean trip speed (in kmph) | Mean 20th percentile speed (in kmph) | Mean 80th percentile speed (in kmph) | Mean 90th percentile speed (in kmph) |
|---|---|---|---|---|---|
| 'Strong and Fearless' or 'Enthused and confident' | 109 | 17.9 | 7.98 | 26.38 | 31.01 |
| Interested but concerned | 562 | 16.4 | 6.72 | 24.22 | 27.83 |

# Appendix D Spatial variation of bikeable network density and off-road bike path and protected bike lane density

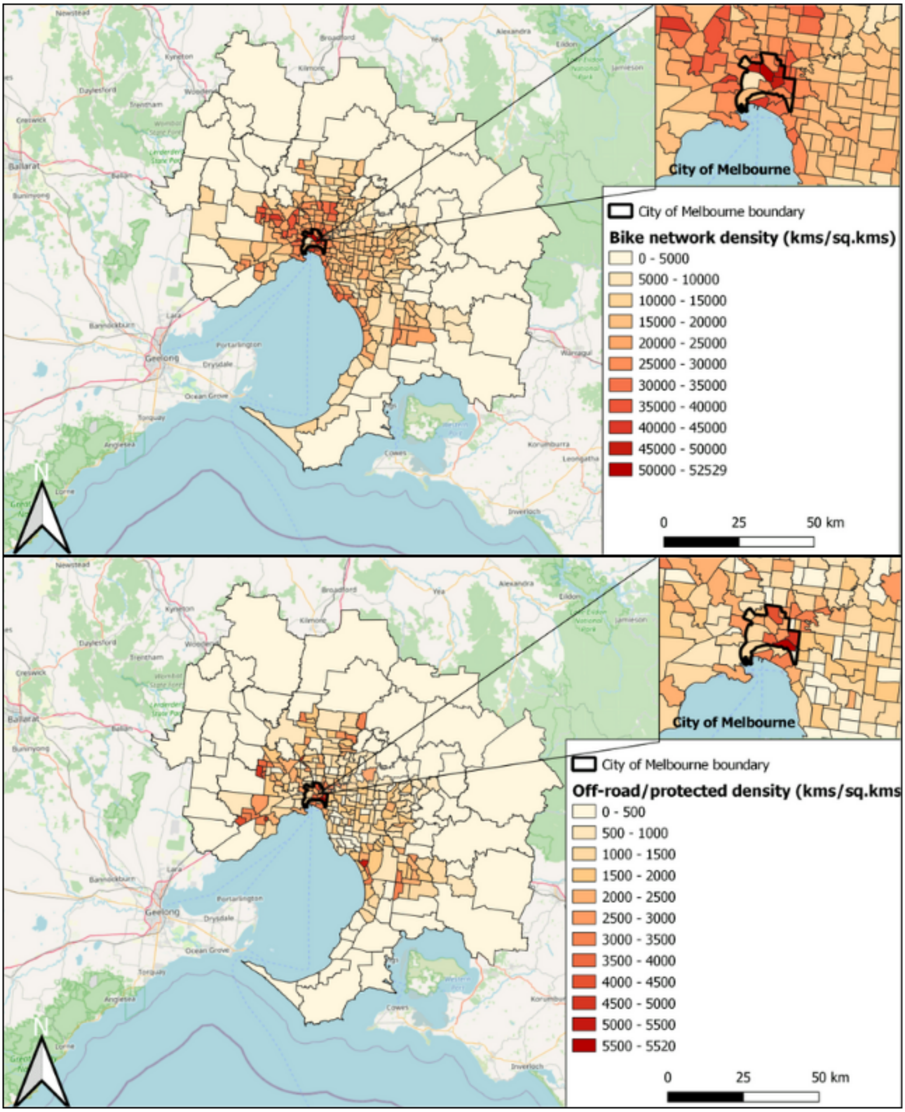



**Figure D2**: Spatial variation of bikeable network density (top) and off-road bike path and protected bike lane density (bottom) across Greater Melbourne (Statistical Area 2 - 2016)

# Appendix E OpenStreetMap tags and values used for extracting bicycle network graph

**Table E5**: OpenStreetMap tags and values used for extracting bicycle network graph

| | | **OpenStreetMap tags** | | | |
|---|---|---|---|---|---|
| | | highway | access | bicycle | area |
| Graph 1 | Included values | *cycleway*<br>*trunk*<br>*primary*<br>*primary_link*<br>*secondary*<br>*secondary_link*<br>*tertiary*<br>*tertiary_link*<br>*residential*<br>*living_street*<br>*service*<br>*trailhead*<br>*unclassified* | | | |
| | Excluded values | | *no*<br>*private* | *no* | *yes* |
| Graph 2 | Included values | *footway*<br>*pedestrian*<br>*path* | | *yes*<br>*designated*<br>*dismount* | |
| | Excluded values | | | | *yes* |