Computers, Environment and Urban Systems 85 (2021) 101553

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

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

Characterizing cycling traffic fluency using big mobile activity tracking data

Anna Brauer [a,b,*], Ville Mäkinen [a], Juha Oksanen [a]

[a] Finnish Geospatial Research Institute FGI, National Land Survey of Finland, 02340 Masala, Finland
[b] Dresden University of Technology, 01069 Dresden, Germany

A B S T R A C T

Mobile activity tracking data, i.e. data collected by mobile applications that enable activity tracking based on the use of the Global Navigation Satellite Systems (GNSS), contains information on cycling in urban areas at an unprecedented spatial and temporal extent and resolution. It can be a valuable source of information about the quality of bicycling in the city. Required is a notion of quality that is derivable from plain GNSS trajectories.

In this article, we quantify urban cycling quality by estimating the fluency of cycling traffic using a large set of GNSS trajectories recorded with a mobile tracking application. Earlier studies have shown that cyclists prefer to travel continuously and without halting, i.e. fluently. Our method extracts trajectory properties that describe the stopping behaviour and dynamics of cyclists. It aggregates these properties to segments of a street network and combines them in a descriptive index. The suitability of the data to describe the cyclists' behaviour with street-level detail is evaluated by comparison with various data from independent sources.

Our approach to characterizing cycling traffic fluency offers a novel view on the cyclability of a city that could be valuable for urban planners, application providers, and cyclists alike. We find clear indications for the data's ability to estimate characteristics of city cycling quality correctly, despite behaviour patterns of cyclists not caused by external circumstances and the data's inherent bias. The proposed quality measure is adaptable for different applications, e.g. as an infrastructure quality measure or as a routing criterion.

1. Introduction

Seeking sustainable, eco-friendly transport alternatives for ever-growing urban areas, local authorities and national agencies have long recognized the potential of cycling. Many countries have implemented strategies to promote cycling and turn it into a safer and more convenient mode of travel (e.g. Bundesministerium für Verkehr, Innovation und Technologie, 2017; Commonwealth of Australia, 2018; Pucher & Buehler, 2012). Raising the modal share of cycling effectively and cost-efficiently requires a solid understanding of the determinants that influence cycling in urban areas. A key enabler is travel behaviour data, which is traditionally collected through questionnaire surveys and manual route logs (Griffin, Nordback, Götschi, Stolz, & Kothuri, 2014). The recruitment of volunteers and the evaluation of these studies are costly and time-consuming even for small amounts of data with little detail and low accuracy (Wang, He, & Leung, 2018). In 2007, the first GNSS-based study on the travel behaviour of cyclists was published (Harvey & Krizek, 2007). Although satellite positioning techniques made recording routes taken by cyclists much more convenient, finding study participants still remained a challenge. Most GNSS-enabled studies face the drawbacks of a small sample size, short data collection periods, and data that quickly becomes out of date (Shen & Stopher, 2014).

Entirely new possibilities opened up when smartphones with built-in GNSS sensors emerged on the market. Research initiatives using custom-designed mobile applications were launched that broadened the range of participants significantly (e.g. Hood, Sall, & Charlton, 2011; Reddy et al., 2010). Even more comprehensive data can be harnessed by repurposing data that is collected by commercial applications (Romanillos, Zaltz Austwick, Ettema, & De Kruijf, 2016). Answering the demand for intelligent ways to keep track of personal fitness and training, companies have developed activity tracking applications, e.g. Strava,[1] Sports Tracker,[2] or Endomondo.[3] Today, the most popular providers have tens of millions of users who utilize the applications to monitor their training and share their activities, predominantly running and cycling, with the community (e.g. Strava, 2018). The data created by cyclists using mobile activity tracking applications comprises trajectories, i.e., sequences of timestamped GNSS measurements, pictures,

* Corresponding author at: **Finnish Geospatial Research Institute FGI, National Land Survey of Finland, 02340 Masala, Finland.**
E-mail addresses: anna.brauer@nls.fi (A. Brauer), ville.p.makinen@nls.fi (V. Mäkinen), juha.oksanen@nls.fi (J. Oksanen).

[1] https://www.strava.com/
[2] https://www.sports-tracker.com/
[3] https://www.endomondo.com/

https://doi.org/10.1016/j.compenvurbsys.2020.101553
Received 16 December 2019; Received in revised form 25 September 2020; Accepted 30 September 2020
Available online 16 October 2020

messages, and rich personal information on fitness and health. This adds up to a huge, rapidly growing set of structurally diverse data, commonly associated with the term *big data* (De Mauro, Greco, & Grimaldi, 2016). The data is collected by a larger user base and with a much wider spatial and temporal coverage than any data collected exclusively for research. A thorough analysis of such data can reveal large-scale spatio-temporal and societal patterns that are especially valuable in the context of urban planning (Romanillos et al., 2016).

Big data is linked to a number of challenges, e.g. concerning the data volume, its structural variety or fast rate of creation (Gandomi & Haider, 2015). Furthermore, it is inherently prone to factors that compromise its veracity, i.e. bias, noise, and uncertainty (Rubin & Lukoianova, 2013).

Uncertainty in mobile tracking data stems largely from GNSS noise and other positioning errors, but the data also suffers from several other issues limiting its usability, including self-selection bias. The user community of mobile tracking applications neither represents a city's population as a whole, nor the subpopulation that use bicycles as a means of travel (Smith, 2015). Cyclists are a highly heterogeneous group that varies demographically and with respect to experience and confidence about cycling (Damant-Sirois, Grimsrud, & El-Geneidy, 2014). Confident cyclists, whose interest in cycling is so high that they have decided to track their activities, tend to be overrepresented in the group of active users of tracking applications (Strava metro data analysis summary, *2018*). Bias can also arise with regard to the purpose and motivation of the recorded cycling trips. Some cyclists monitor only training sessions, while others track their commuting and other utilitarian trips or leisurely recreational activities (Bergman & Oksanen, 2016b).

Moreover, mobile tracking data is personal, potentially sensitive data. Therefore, the protection of the cyclists' privacy must be prioritised, often at the expense of data utility (Primault, Boutet, Mokhtar, & Brunie, 2018).

Despite these aspects, big mobile tracking data can be harnessed to obtain information on the cyclability of cities and urban areas, i.e. the quality and distribution of suitable bicycling infrastructure. To communicate this information, we require measures that support the identification of spatio-temporal cycling patterns and facilitate the comparison of cycling on different streets or in different neighbourhoods. *Traffic fluency*, i.e. the smoothness of the traffic flow, is a well-known concept for motorized traffic. The degree of fluency or its opposite, congestion, is usually determined by measuring the speed of vehicles, travel time, or traffic volumes (Rao & Rao, 2012). In this article, we reinterpret fluency as an attribute of the cycling traffic. Like vehicles in uncongested traffic, cyclists travel *fluently* if their motion is steady and continuous, and if they are free to cycle at a comfortably fast pace without being forced to brake or halt. In previous studies, researchers have found that the majority of cyclists favour continuous infrastructure with an even surface that is segregated from other road users (Caulfield, Brick, & McCarthy, 2012; Sener, Eluru, & Bhat, 2009; Stinson & Bhat, 2005). They strongly dislike stopping and waiting (Menghini, Carrasco, Schüssler, & Axhausen, 2010; Stinson & Bhat, 2005). In this sense, the idea of fluent cycling corresponds well to cyclists' preferences.

In this work, we present an approach to estimating the quality of urban cycling using big mobile tracking data. Our method extracts properties characterizing the fluency of cycling traffic from a large set of cycling trajectories. By aggregating them to segments of a street network, we obtain quantities that describe the movement and stopping behaviour of cyclists on each segment. With the definition of a cycling traffic fluency index, we show one possibility of combining these normalized quantities into a single quality measure that facilitates visualization. To evaluate the veracity of the derived data, i.e. its representativeness and correspondence to real-world circumstances, we compare it to traffic light data, trajectories recorded by a volunteer, and data obtained in a field study.

The article is structured as follows. First, we give an overview of related studies that utilize a large set of cycling trajectories. We then introduce our data and methods for 1) trajectory processing and 2) veracity evaluation. Finally, we review and discuss the results.

2. Related work

The first efforts to utilize crowdsourced cycling trajectories were made when mobile tracking applications were still in their early development. Addressing the challenges associated with large volumes of trajectories, Schüssler and Axhausen (2009) published a processing procedure for raw GNSS trajectories that are unaccompanied by further background information. The origin of their test data, however, was a study carried out by a private sector company to study the placement of billboards. Subsequently, Menghini et al. (2010) showed that it is possible to estimate a route choice model for cyclists from precisely the same data. They noted that the absence of socio-demographics and the involvement of participation inequality were limitations of the data.

Authorities in different countries developed their own applications to analyse the behaviour of local cyclists or promote cycling by providing benefits to frequent riders. Although these applications have a potentially smaller user base, they can be tailored to gather additional data that is valuable for research, e.g. the trip purpose. Examples include the studies by Hood et al. (2011) and Dane, Feng, Luub, and Arentze (2019). Both estimate a route choice model for bicycles or e-bikes, respectively.

The number of researchers who aim to create value from GNSS cycling trajectories collected solely for non-research purposes by commercial tracking applications has been growing in recent years. The studies cover a wide range of application areas, yet one recurring theme is popularity. Ferrari and Mamei (2013), for example, use kernel density estimation to reveal the most popular locations for different sports. As a measure of the cyclability of a city, they also propose an index that reflects the correlation of cycling routes with mobile activity tracking data. Oksanen, Bergman, Sainio, and Westerholm (2015) show that privacy-preserving heat maps can be generated from crowdsourced GNSS cycling trajectories, thus providing a way to visually communicate the popularity of different infrastructure with cyclists. Subsequently, Bergman and Oksanen (2016a) present an approach to utilize the data for popularity-based routing. Similarly, Baker et al. (2017) developed a process to model the appreciation of roads in a network as a way to improve routing for cyclists. Using tracks obtained from the route-sharing platform GPSies, Sultan, Ben-Haim, Haunert, and Dalyot (2015) analyse the usage share of different types of infrastructure.

While all of the previously mentioned research utilizes raw trajectories, acquiring this type of data is difficult, as application providers need to be wary of privacy concerns. Strava recognized the possibility to sell their data in an aggregated form that is adjusted towards representativeness (Strava, 2019). Recent research shows that the data provided by this service can be utilized to monitor bicycle traffic volumes (Griffin & Jiao, 2015) and how the cycling traffic flow reacts to infrastructure changes (Boss, Nelson, Winters, & Ferster, 2018). Additionally, it can be used to reveal the impact of determinants such as demographic factors or infrastructure characteristics (Hochmair, Bardin, & Ahmouda, 2019).

To the best of our knowledge, this study is the first to derive properties of the dynamics of city cycling from a large set of GNSS trajectories to form a measure for the cyclability of a city.

3. Data

The primary data of this work consists of 50,357 GNSS trajectories from 3694 cyclists travelling in the Helsinki metropolitan area (Helsinki, Espoo, Vantaa, and Kauniainen). The trajectories were recorded with a mobile sports tracking application between 2010 and 2012 and were made public by the application users. Each trajectory is associated with a cyclist pseudo id, which allows us to identify trajectories recorded by the same cyclist. The sampling rate of the trajectories is consistently high

(mean 1.45 s).

The dataset is biased in several ways. We can observe participation inequality since only 10% of the cyclists account for 65% of the activities and 67% of the total cycled distance (Fig. 1). Most activities, 77%, were recorded between May and September. The temporal variation of the recording suggests that the dataset contains commuting trips as well as leisure cycling activities (Fig. 2).

Any additional data used for our analyses is openly available. To adjust the trajectories, we require street network data which then serves as a target for aggregating trajectory properties. We utilized the street network data made available by OpenStreetMap (OSM).[4] We only included features that are traversable by cyclists. Since the length of the features varies considerably, we split them into approximately 25-m-long segments. This way, we obtain uniform features as base elements for the aggregation. With a segment length of 25 m, the level of detail is as high as possible while guaranteeing a minimum number of two GNSS measurements per trajectory and segment in most cases.

To evaluate the results, we used traffic-light data retrieved from the Helsinki Region Infoshare service.[5] The data does not contain the exact position of the traffic lights, but it is rather a set of point features marking intersections controlled by traffic lights.

Furthermore, we examined 47 locations in central Helsinki for factors that could potentially obstruct cycling. The locations were not chosen randomly, but in accordance with initial results obtained from the trajectory dataset.

4. Method

We designed a process (Fig. 3) that takes a set of GNSS trajectories as input, processes them and extracts properties related to cycling traffic fluency (CTF). These properties are aggregated to the road network of the study region and finally combined into a measure for CTF. Each processing step is described in detail in the following. The second part of this section deals with our methodology for validating the veracity of the derived properties and analysing the results of the CTF estimation.

4.1. Trajectory processing

4.1.1. Trajectory smoothing

To reduce the GNSS noise, we executed kernel-based trajectory smoothing with a Gaussian kernel function (Schüssler & Axhausen, 2008). For each point z_i in the trajectory, both dimensions of its smoothed counterpart s_i are calculated as

$$s_i(l) = \frac{\sum_{j=i-N}^{i+N} w_{i,j} z_j(l)}{\sum_{j=i-N}^{i+N} w_{i,j}}, \tag{1}$$

where $l \in x, y$. The weight factors $w_{i,j}$ are calculated using the Gaussian function

$$w_{i,j} = w(\Delta t_{i,j}) = exp\left(-\frac{\Delta t_{i,j}^2}{2\sigma^2}\right), \tag{2}$$

where $\Delta t_{i,j}$ is the time difference between the points z_i and z_j. Due to the low frequency of outliers in the trajectories, we opted for a parameter combination that resulted in mild smoothing and preserved sharp turns as much as possible ($N = 2$, $\sigma = 1.2$).

4.1.2. Map matching

The smoothed trajectories were map-matched to the street network. We utilized a map matching procedure that is based on Hidden Markov Models (Newson & Krumm, 2009). For every smoothed trajectory point s_i, the procedure estimates the emission probabilities for every nearby street segment. The probability that the point s_i was recorded on street segment r_j is calculated as

$$p(s_i|r_j) = \frac{1}{\sqrt{2\pi}\sigma_z} exp\left(-\frac{1}{2}\frac{\|s_i - x_{i,j}\|^2}{\sigma_z^2}\right), \tag{3}$$

where $x_{i,j}$ is the closest point to s_i on the road segment r_j. Furthermore, the algorithm calculates transition probabilities. The transition probability $p(x_{i+1,n}|x_{i,m})$ is the probability that a smoothed point s_{i+1} corresponds to a map-matched point $x_{i+1,n}$ on road segment r_n if the previous point s_i corresponds to a map-matched point $x_{i,m}$ on road segment r_m. For correctly matched pairs of points, the Euclidean distance of the points s_{i+1} and s_i should be relatively close to the distance along the road segments between points $x_{i+1,n}$ and $x_{i,m}$. Therefore, the transition probabilities are calculated as

$$p(x_{i+1,n}|x_{i,m}) = \frac{1}{\beta} exp\left(-\frac{\left|\left\|x_{i+1,n} - x_{i,m}\right\|_{\text{road}} - \|s_{i+1} - s_i\|\right|}{\beta}\right) \tag{4}$$

With these two sets of probabilities, the optimal sequence of map-matched points can be calculated using the Viterbi algorithm (Forney Jr., 1973).

4.1.3. Stop detection

According to Spaccapietra et al. (2008), a trajectory can be divided into alternating sequences of stops and moves, i.e. sequences of trajectory points where the cyclist either remains in one place or travels, respectively. Since cycling traffic fluency manifests in both trajectory components, we considered stop- and movement-related properties of trajectories. We identified stops with a spatio-temporal density-based clustering algorithm (CB-SMot; Palma, Bogorny, Kuijpers, & Alvares, 2008). A stop is defined as a sequence of trajectory points within a neighbourhood of radius *Eps* that lasts at least *min_time* seconds. We choose *Eps* dynamically for each trajectory as the mean Euclidean distance between consecutive points. The *min_time* was set to 10 to minimize the chance of detecting false positives. Each stop identified by the CB-SMoT algorithm was mapped to a street network segment by majority vote of the map-matched counterparts of the points that belong to the stop. For simplicity, a stop can be represented by its centroid, which is the arithmetic mean of the points. We also determined the stop duration, i.e. the time that passes between the first and the last point of the stop.

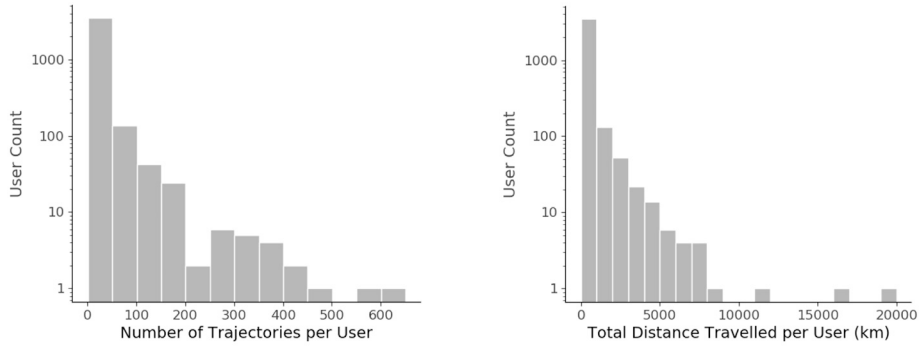
4.1.4. Extraction of movement-related properties

For each trajectory point, we initially obtained two movement-related properties: speed and acceleration. The speed of a trajectory point was calculated as the average of the speed based on the time intervals and the distances between the map-matched representations of the point and its successor and predecessor. The acceleration was calculated similarly, using speed instead of distance. Subsequently, we divided each trajectory into short, consecutive point sequences so that the points in one sequence were matched to the same street segment. We refer to these partial trajectories as *runs*, denoted by χ (Fig. 4). By comparing the orientation of the segment to the direction of travel of the trajectory, we ensured that a trajectory passing a perpendicular street did not create a run for a perpendicular segment.

We denote the speed and acceleration of a run χ by speed(χ) and acceleration(χ). These run properties were calculated as the average speed and acceleration of the map-matched points belonging to run χ. If the run consists of only one map-matched point, speed(χ) and acceleration(χ) equal the speed and acceleration of the point. Consequently,

[4] https://www.openstreetmap.org/

[5] https://hri.fi/data/en_GB/dataset/helsingin-espoon-ja-vantaan-liikennevaloristeykset



(a) Number of cycling trajectories recorded per user. (b) Total distance travelled per user.

Fig. 1. (a) Number of recorded cycling trips and (b) sum of travelled kilometers per application user.

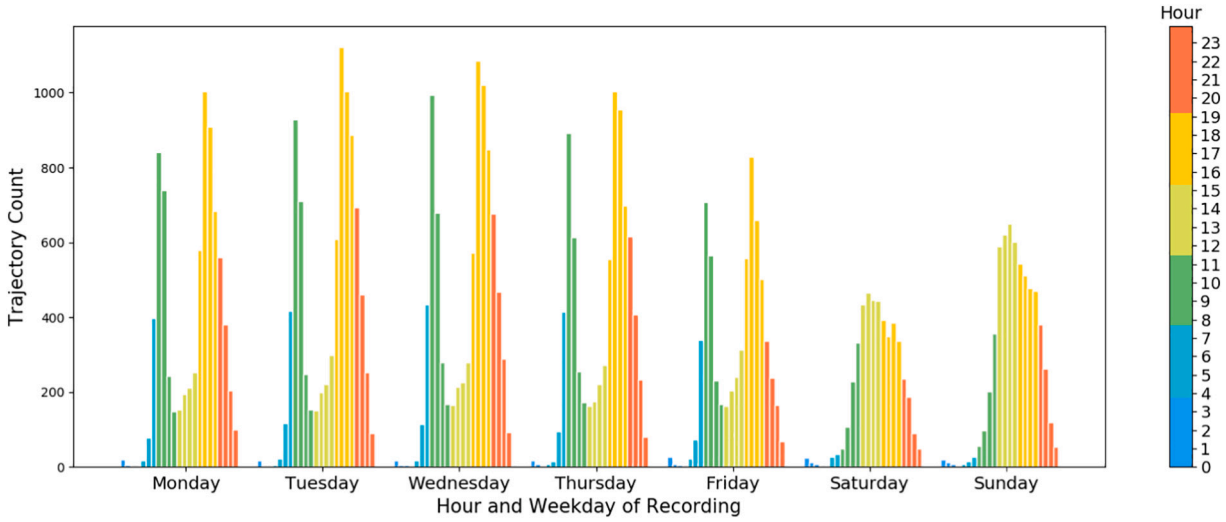


Fig. 2. Number of trajectories per hour and day of the week.

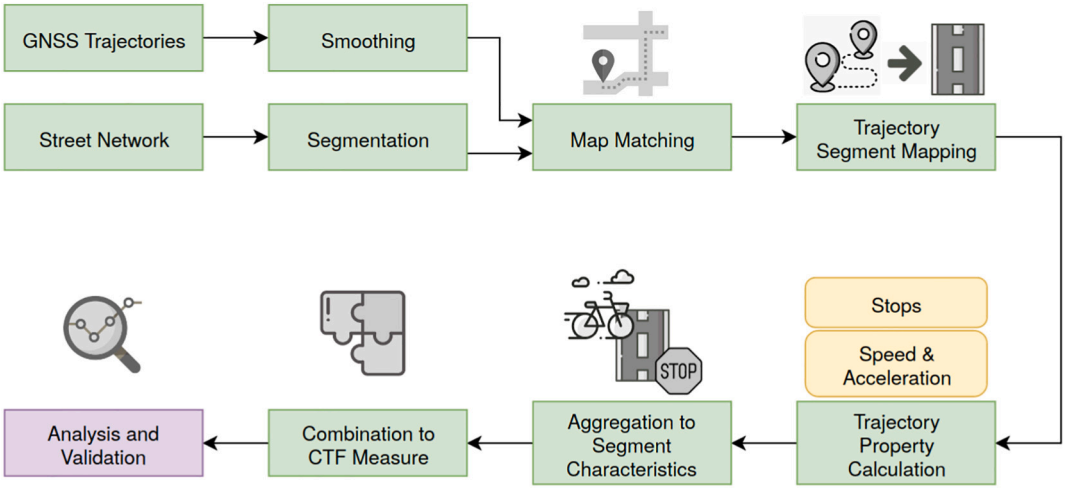


Fig. 3. Overview of the work process.

length and duration of a run are defined as:

- length(χ): the length of the street segment to which the run is mapped;
- duration(χ): the time needed to travel the street segment at speed(χ).

To exclude outliers, runs with an unrealistic value for speed(χ) or acceleration(χ) and runs at the very beginning or end of a trajectory were discarded. Ultimately, the set of runs X represents the original trajectories but cannot be used to reproduce them completely.

Using a subset of runs $X_t \subset X$, i.e. the set of runs of a trajectory t that do not contain stop points, we define the mean travelling speed of t as

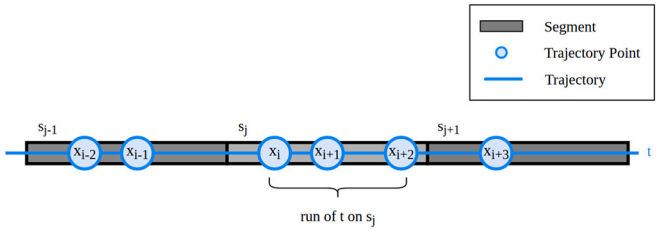


Fig. 4. A *run* is a sequence of consecutive trajectory points that are matched to the same segment.

$$\bar{v}_t = \frac{\sum_{\chi \in X_t} \text{length}(\chi)}{\sum_{\chi \in X_t} \text{duration}(\chi)} \quad . \tag{5}$$

With this definition, the parts of the trajectories that are classified as stops do not affect the mean travelling speed.

Finally, we estimated whether a segment was traversed faster or slower in comparison to the trajectory's mean travelling speed by calculating the speed ratio of a run χ:

$$\text{speed\_ratio}(\chi) = \frac{\text{speed}(\chi)}{\bar{v}_t}, \tag{6}$$

where t corresponds to the trajectory of which χ is a part.

4.2. Aggregation of street network characteristics

The extracted segment-specific properties were only then aggregated to a segment of the street network if data from at least 10 cyclists was available. This density threshold ensures a basic level of trajectory diversity, which increases the likelihood that the aggregated characteristics are representative. Additionally, the threshold increases the protection of the application users' privacy.

The stops derived from individual trajectories were straightforwardly aggregated to the street segments. First, we defined X_s as the set of runs that were mapped to a specific segment. The index s refers not only to the segment, but also to the direction of travel. Thus, only trajectories that traverse the segment in the same direction contribute to the same values. We then calculated the number of runs $|X_s|$ on the segment, the number of stops C_s that were assigned to the segment, and the average duration T_s of these stops. Another important quantity is the ratio of cyclists who stopped on the segment. We refer to this as the stop ratio $\widehat{C}_s$:

$$\widehat{C}_s = \frac{C_s}{|X_s|} \quad . \tag{7}$$

To aggregate the movement-related trajectory properties, we calculated the segment-wise average speed v_{seg}, acceleration a_{seg}, and speed ratio $\widehat{v}_{\text{seg}}$ for each street segment s:

$$v_{\text{seg}}(s) = \frac{1}{|X_s|} \sum_{\chi \in X_s} \text{speed}(\chi) \quad , \tag{8}$$

$$a_{\text{seg}}(s) = \frac{1}{|X_s|} \sum_{\chi \in X_s} \text{acceleration}(\chi) \quad , \tag{9}$$

$$\widehat{v}_{\text{seg}}(s) = \frac{1}{|X_s|} \sum_{\chi \in X_s} \text{speed\_ratio}(\chi) \quad . \tag{10}$$

At the end of the aggregation phase, we have six properties for each street segment that is traversed by ten or more trajectories: the number of stops C_s, the average duration of the stops T_s, the stop ratio $\widehat{C}_s$, the average speed $v_{\text{seg}}(s)$, the average acceleration $v_{\text{acc}}(s)$, and the speed ratio $\widehat{v}_{\text{seg}}(s)$.

4.3. Combination into descriptive indices

The possibilities of transforming the segment characteristics and combining them in a single cycling quality measure are manifold. In the following, we present a variant that is designed to facilitate visual analyses of the results.

All the characteristics were at first transformed into normalized indices. Starting with the movement-related characteristics, we converted the speed ratio into the speed ratio index I_{speed}:

$$I_{\text{speed}}(s) = min\left(1, \frac{1}{2} + \sqrt[3]{\frac{\widehat{v}_{\text{seg}}(s) - 1}{10}}\right) \quad . \tag{11}$$

This formula emphasizes the relation of the speed ratio of segment s to 1, i.e. the mean travelling speed (Fig. 5a). This emphasis is sensible because for most segments, the speed ratio tends to be close to 1. The index has a strong linear dependence on $\widehat{v}_{\text{seg}}(s)$ near the value 1, but the dependence weakens as the difference grows.

The acceleration of a segment is more difficult to interpret. We converted the acceleration values into the acceleration index I_{acc} with

$$I_{\text{acc}}(s) = \begin{cases} exp\left(-a_{\text{seg}}(s)\right) & a_{\text{seg}}(s) > 0\frac{m}{s^2} \\ exp\left(2.5a_{\text{seg}}(s)\right) & \text{else} \quad . \end{cases} \tag{12}$$

According to this definition, all changes of the travel speed are considered unwanted, but deceleration is penalized more than positive acceleration (Fig. 5b).

We combined the two indices into a single measure, referred to as the movement-related index I_{move}, using the harmonic mean:

$$I_{\text{move}} = 2\frac{I_{\text{speed}} \cdot I_{\text{acc}}}{I_{\text{speed}} + I_{\text{acc}}} \quad . \tag{13}$$

The harmonic mean guarantees that I_{move} can reach high values only if both the constituents also have high values. This corresponds to moving continuously at a steady, above-average speed.

Similarly, we converted two stop-related characteristics, i.e. the average duration T_s of the stops on a segment and the ratio of cyclists $\widehat{C}_s$ who stop on a segment s, into corresponding indices. The average stop duration was converted into the mean stop duration index:

$$I_{\text{stop}}(s) = \begin{cases} 1 & T_s < 10 \\ 0.8 & 10s \le T_s < 15 \\ 0.6 & 15s \le T_s < 20 \\ 0.4 & 20s \le T_s < 25 \\ 0.2 & 25s \le T_s < 30 \\ 0.01 & T_s \ge 30 \end{cases} \tag{14}$$

The index classifies the segments into six categories. The highest class has the value 1, which means that segments with no significant stops are not penalized at all. Average stop duration values longer than 30 s receive the greatest penalty.

The ratio of cyclists who had to stop on a segment was classified to form the stop ratio index:

$$I_{\text{stop\%}}(s) = \begin{cases} 1 & \widehat{C}_s < 0.01 \\ 0.8 & 0.01 \le \widehat{C}_s < 0.05 \\ 0.6 & 0.05 \le \widehat{C}_s < 0.1 \\ 0.4 & 0.1 \le \widehat{C}_s < 0.2 \\ 0.2 & 0.2 \le \widehat{C}_s < 0.3 \\ 0.01 & \widehat{C}_s \ge 0.3 \end{cases} \tag{15}$$

Segments where the percentage of stopping cyclists is below 1% are assigned the highest possible value 1. If the percentage is higher than 30%, the segments are rated in the lowest category.

The category choices for both stop-related indices were guided by

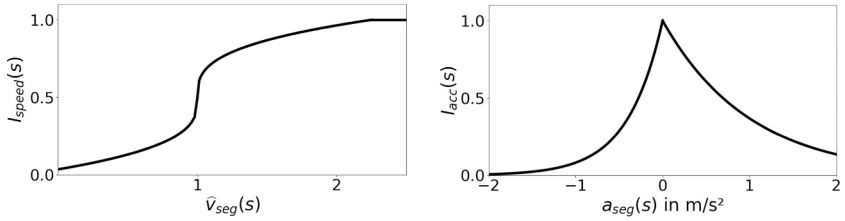


Fig. 5. Index transformation functions that normalize (a) the speed ratio and (b) the acceleration of a segment *s*.

inspecting the distribution of the values derived from the dataset. Therefore, the indices may require adjustment if they are applied to other datasets.

In contrast to the I_{move} index introduced previously, we combined the two stop-related indices using the arithmetic mean:

$$I_{\text{stop}} = \frac{I_{\text{stop}} + I_{\text{stop\%}}}{2} \quad . \tag{16}$$

The rationale behind this decision was that the combined index should not receive very low values if only one of the constituents is low. In other words, even if stopping is guaranteed when passing through a segment, it is not penalized heavily if the average stop duration is short, and vice versa, even a long average stop duration cannot lower the index value too much if stops are extremely rare.

With these definitions, we specified the cycling traffic fluency index (the CTF index) as the final quality measure:

$$I_{\text{fluency}} = (1+\beta)\frac{I_{\text{move}} \cdot I_{\text{stop}}}{\beta \cdot I_{\text{move}} + I_{\text{stop}}} \quad , \tag{17}$$

where the factor β balances the relative weight of the two index components. Again, we used the harmonic mean because both constituents need to be high to consider the dynamics on a street segment fluent. This way, fluency hindrances indicated by speed, acceleration, or both of the stop-related properties translate directly into the CTF index. Fig. 6 shows the behaviour of the index and the implications of the choice of the mean function.

4.4. Validation of the derived data

In pursuit of knowledge about the veracity of the derived data, we turned to reference data of different kinds and origins. To validate individual stops, we analysed their distance to traffic lights and intersections of the street network.

Furthermore, we identified stop hot spots, i.e. locations where stops accumulate, to find patterns in the set of detected stops. The hot spots were constructed by gathering the centroids of the stops into clusters using DBSCAN (Ester, Kriegel, Sander, & Xu, 1996). We defined the minimum number of stops in a cluster as 10 to ensure a certain level of significance. The stop duration of a hot spot corresponds to the average duration of all stops in the cluster. To calculate the stop ratio of a hot spot, we identified all the street segments that intersected the buffered convex hull of all the stop centroids in the cluster. We then divided the number of stops in the cluster by the number of trajectories that passed any of the segments associated with the hot spot. The buffer around the convex hull increased the noise tolerance. A buffer width of 3 m was experimentally determined to be sufficient for our data.

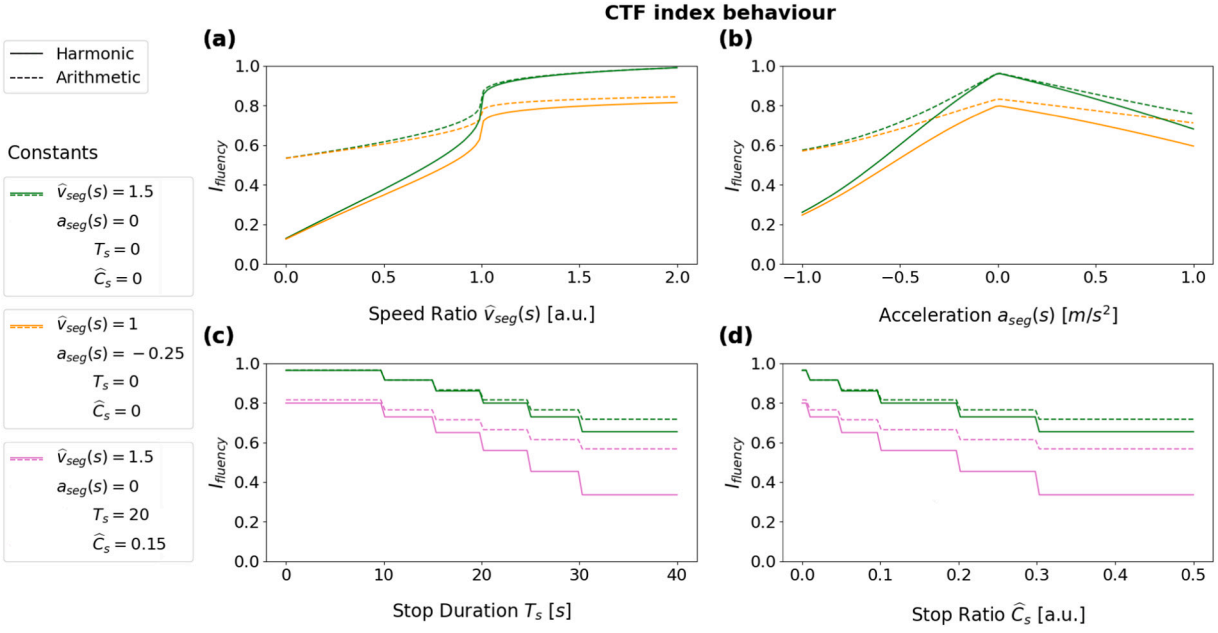


Fig. 6. Dependency of the CTF index on its input variables and the impact of the choice of the mean (arithmetic or harmonic) in the final index composition step (Eq. 17) with $\beta = 1$. The variables which are not displayed are assigned constant values that correspond to almost optimal conditions (green) or hampered cycling indicated by either the movement-related characteristics (orange) or the stop-related characteristics (violet). If the harmonic mean is used, a single movement-related variable that indicates unfavourable conditions can affect the outcome more significantly (a and b). The impact of the stop-related characteristics is limited regardless of the chosen mean, unless both indicate significant obstructions (c and d). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

We estimated the cause of the stop hot spots by sorting them into three classes: “traffic light”, “intersection”, and “other”. The classification was based on the distance of the hot spot centroids to the closest traffic light and intersection. It adhered to the following rules:

1. If the distance to the closest traffic light is less than 30 m, classify the hot stop as “traffic light”.
2. Else, if the distance to the closest intersection is less than 15 m, classify as “intersection”.
3. Else, classify as “other”.

Movement-related segment characteristics such as speed and acceleration should, if true to the situation on the street, partially explain the behaviour of individual cyclists. To verify this hypothesis, we utilized 123 trajectories recorded by a volunteer that were not part of the large trajectory set used for aggregation. We compared the speed, acceleration and speed ratio profiles of the test trajectories to profiles generated from the corresponding characteristics of the street segments traversed by the trajectories. More precisely, we created two sequences per test trajectory and property: one contained the property values of all runs in the trajectory, the other the aggregated values of the corresponding segments. The correlation of the two sequences was estimated with Pearson’s correlation coefficient (Rodgers & Nicewander, 1988). If the properties of the test trajectories correlated with the segment characteristics, it would be a clear indicator of the ability of the segment characteristics to reflect the on-street cycling conditions and predict the behaviour of cyclists in view of the built environment.

Complementary to these validation efforts, we carried out a visual analysis and a field study with a focus on particularly interesting features, e.g. junctions, parallel ways, and dedicated cycling facilities.

4.5. Exploratory result analysis

To analyse the results of the CTF estimation, we inspected the variation of the index values with respect to space and time. Previously, we showed how the different indices were calculated using the whole trajectory dataset. To investigate the variation with respect to time, we computed the index values considering only stops and runs that fell into a specific time window. We aggregated the data in hourly intervals to examine the diurnal variation. Similarly, we compared data recorded in the winter months, between December and March, to data recorded during the rest of the year.

On the premise that the CTF varies between different types of cycling infrastructure, we used OSM metadata to group the street segments according to their type of infrastructure. This enabled the comparison of, e.g., cycling on streets and on dedicated cycleways.

One possible application of the CTF index is to use it as a routing criterion. We explored the results of simple CTF-based routing by taking pairs of starting and destination points in the study area and comparing the “most fluent” route with the shortest, the fastest, and the most popular route. We facilitated routing using the Dijkstra shortest path algorithm (Dijkstra, 1959) with different edge weights (Table 1). Since the Dijkstra algorithm works by minimizing the cumulative edge weights, we used the inverse of the number of runs $|X_s|$ to determine the most popular route, and $1 - I_{\text{fluency}}$ for the most fluent one.

Table 1
Dijkstra edge weights of a segment s for different routing criteria. $\|s\|$ denotes the segment length, $\bar{v}$ the average mean segment speed in the study area, X_s the set of runs over segment s, and $\min(I_{\text{fluency}})$ the minimum value of I_{fluency} in the study area.

Route	Edge Weight	
	$I_{\text{fluency}}(s)$ available	$I_{\text{fluency}}(s)$ unavailable
Shortest	$\|s\|$	$\|s\|$
Fastest	$\|s\|/v_{\text{seg}}(s)$	$\|s\|/\bar{v}$
Most popular	$\|s\|/\|X_s\|$	$\|s\|$
Most fluent	$\|s\| \cdot (1 - I_{\text{fluency}}(s))$	$\|s\| \cdot (1 - \min(I_{\text{fluency}}))$

5. Results

Meaningful results can only be obtained if the smoothed, map-matched trajectories ensure a certain quality level. Randomized visual trajectory inspection evinced a sufficiently good quality of the results for all of the pre-processing steps. Map matching led to the most significant changes, shifting the original trajectory points 4.6 m on average. We observed some common mismatching issues, e.g. on parallel lanes, or where cyclists choose alternative paths not included in the OSM street network.

The aggregation of trajectory properties to the street network revealed spatial popularity bias. A few streets and cycleways were cycled so frequently that they are traversed by hundreds of trajectories. The majority of all the street segments passing the density threshold are traversed by only a few tens of trajectories (Table 2).

5.1. Identification of stop hot spots

Our analysis identified 180,606 stops in the trajectories. Half of all the stop centroids were found to be closer than 43 m to the nearest point representation of a traffic light. The real-life distance was probably smaller, considering that the point representations of traffic lights are often located in the middle of an intersection. 85% of all stops appeared closer than 22 m to the nearest intersection in the street network. In general, the stops correlated significantly with both intersections and traffic lights.

We identified 2739 stop hot spots by clustering individual stops. Their locations matched the findings for individual stops, as a large majority of all hot spots coincided with intersections (Fig. 7). Based on the heuristic cause inference, 27% of all the stop hot spots were determined to be traffic light-induced and 64% intersection-induced, which leaves 9 for the third group, where the reason for stopping could not be inferred.

Although not the largest, the group of traffic light-induced hot spots can be considered the most significant. It has the highest average stop duration (22.1 s) and stop ratio (0.13). On average, traffic light-induced hot spots were also derived from the largest number of individual stops (125). This count was much smaller for the group of intersection-induced hot spots (22) or hot spots with an unknown cause (15).

The average stop duration at the stop hot spots was distributed relatively equally over the whole study area and ranged between 10s and 30s for 95% of all hot spots. We observed more variation for the stop ratio. In central Helsinki, the hot spots with especially large stop ratios greater than 0.25 were most often located at large intersections in the city centre, while those with small ratios tended to occur on less busy infrastructure along the shoreline. There were a number of hot spots characterized by a long average stop duration (>30s) or a high stop ratio (>0.20) which stood out because of their curious location. Their cause was unknown and they contained very few stops, not much more than

Table 2
The distribution of the data density in the study region does not show a steady decline for increasing numbers of trajectories. Instead, the number of segments having a high trajectory count is disproportionally high.

Trajectory Count	Percentage of Segments
0	36
1–9	31
10–19	8
20–49	9
50–99	6
100–199	4
>200	6

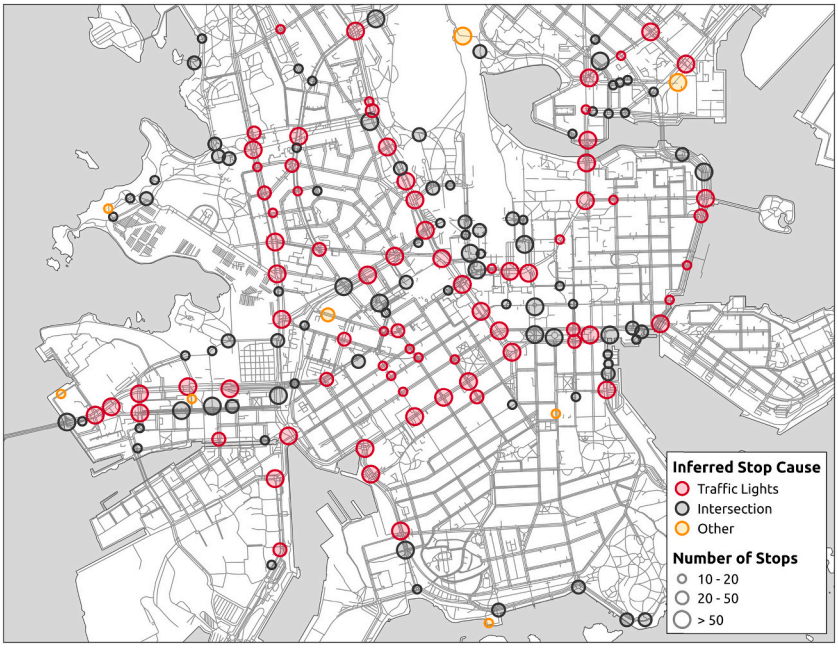


Fig. 7. Inferred stop cause of stop hot spots in central Helsinki in conjunction with the number of stops per hot spot.

the threshold of 10 individual stops.

We conducted a field study in which we visited 47 of these hot spots in central Helsinki. It revealed that the heuristic cause inference identified traffic light-induced hot spots correctly most of the time, but it had problems distinguishing between intersection-induced ones and those of a different origin. Most hot spots labelled as "intersection-induced" were indeed close to some kind of intersection that was, however, most often not the reason for the stopping behaviour inferred from the data. The true reason was not always unambiguously identifiable. Some hot spots were found close to points of interest such as a hospital, a metro station, or a viewpoint. In two cases, we came across stairs in the nearby surroundings. In another two, there was nothing in sight that could explain why cyclists stopped at that particular location.

5.2. *Descriptive analysis of movement characteristics*

Using street network metadata and profound knowledge of the study region, we examined the movement-related segment characteristics. On average, the segment speed in the study region was 6.24 m/s, whereas the speed ratio averaged 1.05, which corresponds to slightly faster cycling than at the mean travelling speed. Segments with low speed and speed ratio values accumulated in the busy city centre of Helsinki. The

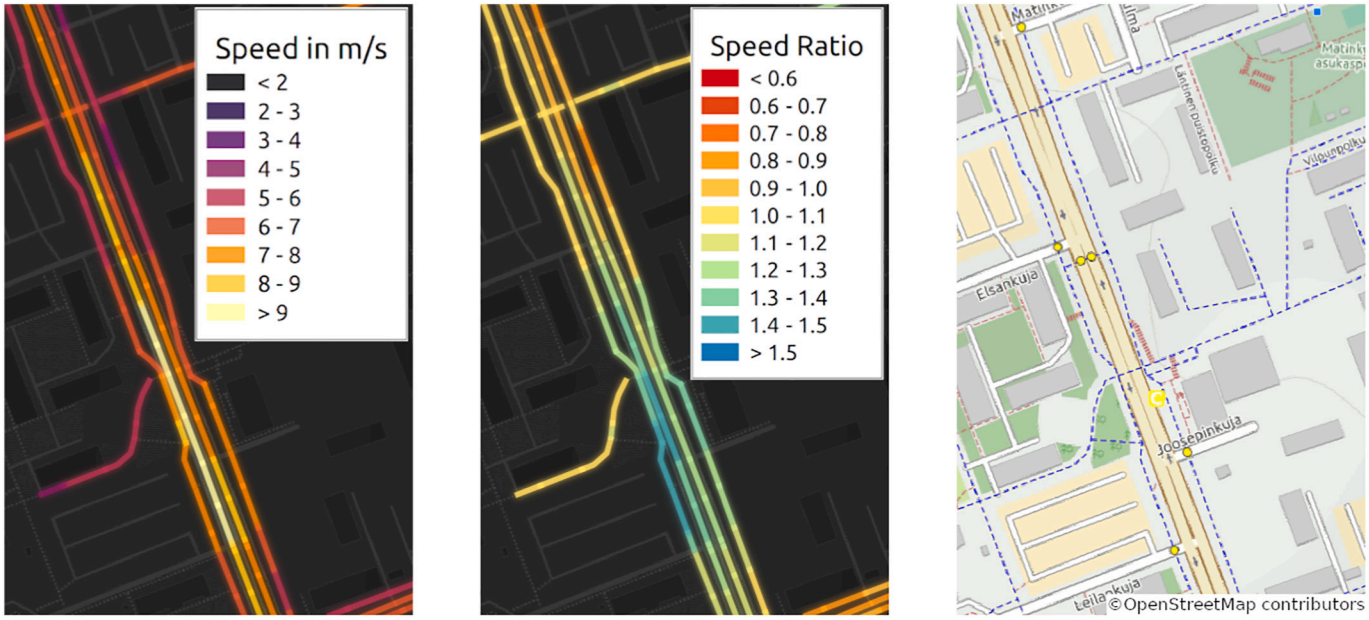


Fig. 8. Speed and speed ratio of cyclists travelling on a street in Espoo. The speed level on the two lanes in the middle is significantly higher than on the surrounding pavement. The speed ratio shows a similar pattern of change for all lanes.

farther from the inner city, the more frequently higher speed values could be observed. Local speed minima corresponded to the location of stop hot spots, and thus to intersections.

With few exceptions, the speed and speed ratio changed gradually between neighbouring segments. Along a single street or path, the segment speed usually did not vary considerably. Sometimes, this continuity was locally interrupted, most often due to an intersection. Parallel ways, e.g. a street with a contiguous pavement, tended to have similar speed ratio profiles. The corresponding absolute speed values, however, could be entirely different (Fig. 8).

In contrast, the acceleration often changed rapidly between neighbouring segments. On infrastructure that allows for continuous cycling, the acceleration fluctuated around zero. The average acceleration of all segments in the study region was 0.04 m/s^2, which was close to zero as well. Where the cycling conditions worsened, extreme acceleration values emerged more frequently, and swift accelerations occurred alongside harsh decelerations. Again, it was mostly intersections that exhibited characteristic patterns of strong accelerating and braking behaviour (Fig. 9). Sharp turns or points of interest, however, had sometimes a similar effect. Segments on rough and narrow paths, as well as ways that are traversed by only a few trajectories, also tended to exhibit extreme acceleration values. It should be noted that the corresponding speed values could suggest continuity even if the acceleration signalled unsteadiness.

5.3. Correlation with individual trajectories

Comparing the properties of an individual trajectory as a function of time to the characteristics of the traversed segments, we found a significant correlation (Fig. 10). One major difference lay in the amplitude of local extrema, which tended to be higher for individual trajectories. As expected, there were also time intervals where the two sequences varied independently, which was especially true for acceleration. These observations were reflected in Pearson's correlation coefficient. On average, Pearson's r equated to 0.62 for the speed and 0.60 for the speed ratio, but only to 0.22 for acceleration. In conclusion, the correlation between the properties of individual trajectories and the characteristics of the corresponding segments was clearly visible, especially for the speed and speed ratio.

5.4. Distribution of the CTF index

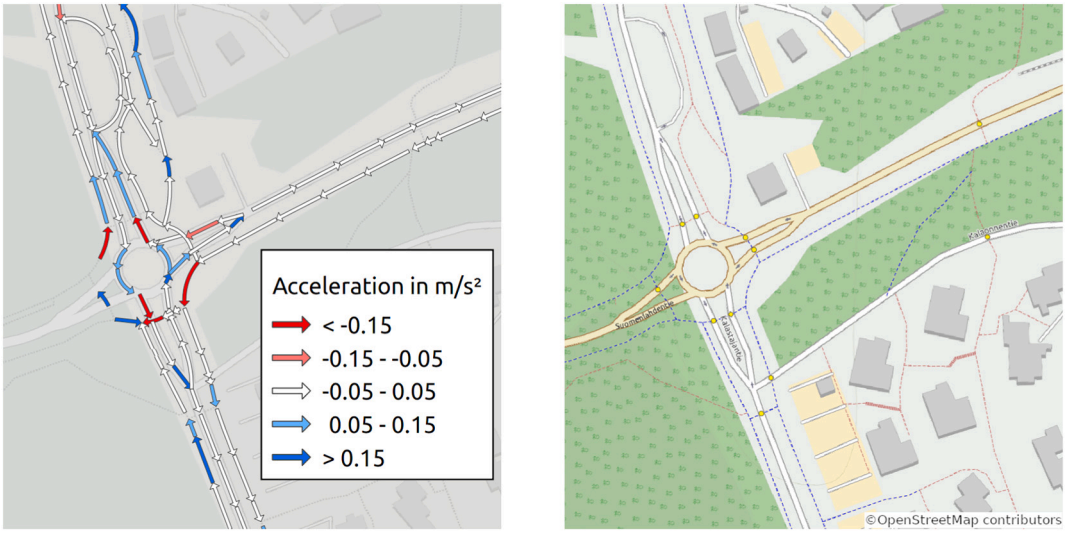
The CTF index I_{fluency} inherits traits of both the movement index I_{move} and the stop index I_{stop}. Since I_{stop} negatively affects only segments that count at least one stop, I_{fluency} equals I_{move} shifted towards 1 for the majority of the segments. The weighting parameter β determines the degree of the shift. In the following, we set $\beta = 1$, weighting both input indices equally. Consequently, some nuances of I_{move} are smoothed out so that I_{stop} can take effect. We note that the optimal choice for β may vary depending on the application.

Our study region was dominated by segments with high (I_{fluency} between 0.7 and 1) and moderate (I_{fluency} between 0.4 and 0.7) CTF index values with a share of 55% and 38%, respectively. Low values ($I_{\text{fluency}} <$ 0.4) accumulated close to intersections (Fig. 11).

Strong variations between segments in the same neighbourhood were much more common for the CTF index than, e.g., for the segment speed v_{seg}. The main reason for this is that the transformation enforced by the speed ratio index I_{speed} emphasizes the difference between the segments' speed ratio and 1. Another factor is the stop index, a combination of two discrete indices. Although segments where the cyclists stopped tended to cluster, they also frequently bordered segments that were not associated with any stop.

Slicing the data into one-hour intervals reduced the number of segments that pass the density threshold considerably. A total of 4680 segments, only 1.4% of all segments in the study area, were used by at least 10 different cyclists every hour between 7 am and 9 pm. The average I_{fluency} per interval varied only about 0.01 between the minimum and maximum. There was a little more variation for the input indices, but I_{acc}, I_{speed}, and the stop-related indices seemed to vary independently of each other. To some extent, the variation appeared to be random. However, considering the CTF towards Helsinki's city centre during the morning rush hour (8–9 am) and noon (12–1 pm), we observed prominent changes. Scattered across the study area, there were spatial clusters of segments showing a distinctive improvement in the CTF between the morning rush hour and noon. With a similar intensity, these changes reversed between noon and the evening rush hour (5–6 pm) (Fig. 12).

Contradicting expectations, the data indicates that the level of CTF was higher in the winter months than in summer. Furthermore, some segments signalled considerably obstructed fluency in the summer, but not in winter. There seems to be no general rule explaining why these



(a) (b)

Fig. 9. Segment accelerations at a roundabout in Espoo. Cyclists either use the street or travel on the pavement and cross the street using zebra crossings. The segments in the intersection area show characteristic patterns of braking and accelerating.

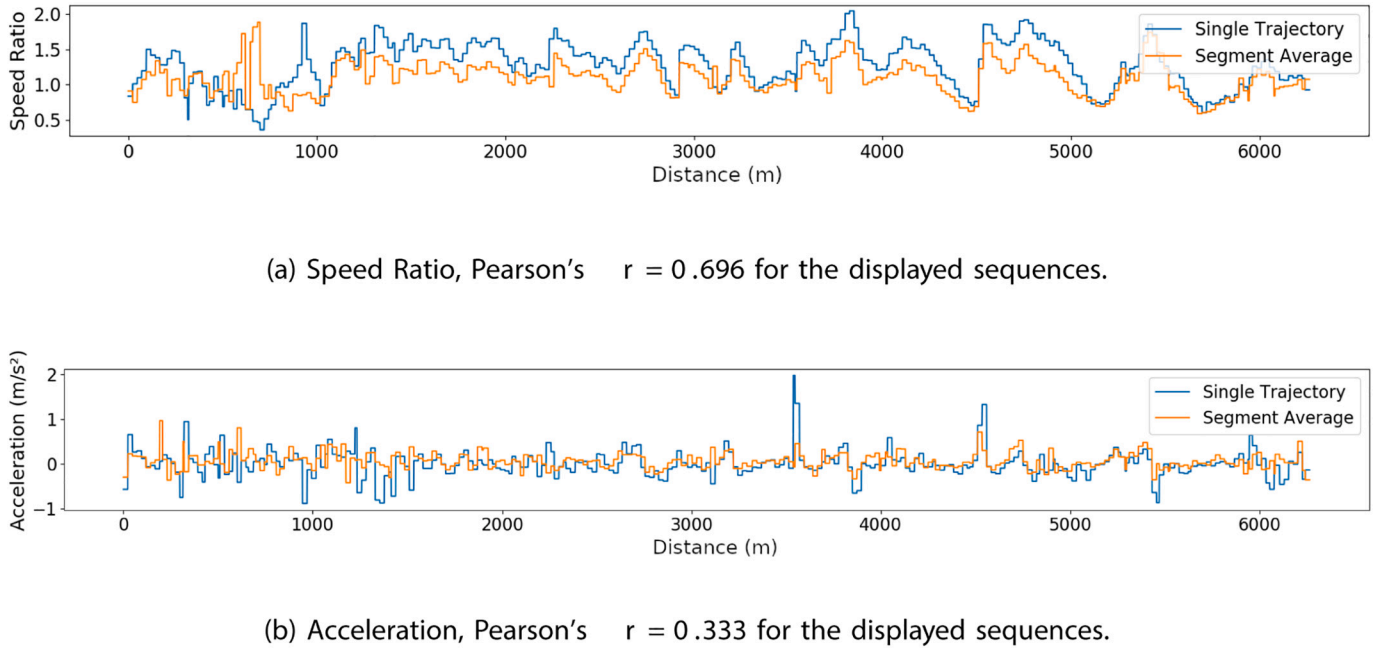


Fig. 10. Speed ratio and acceleration of a trajectory from the test set (blue) in comparison to the characteristics of the passed segments (orange). We observe a weak correlation between the two series for acceleration, and a stronger correlation for speed and speed ratio. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

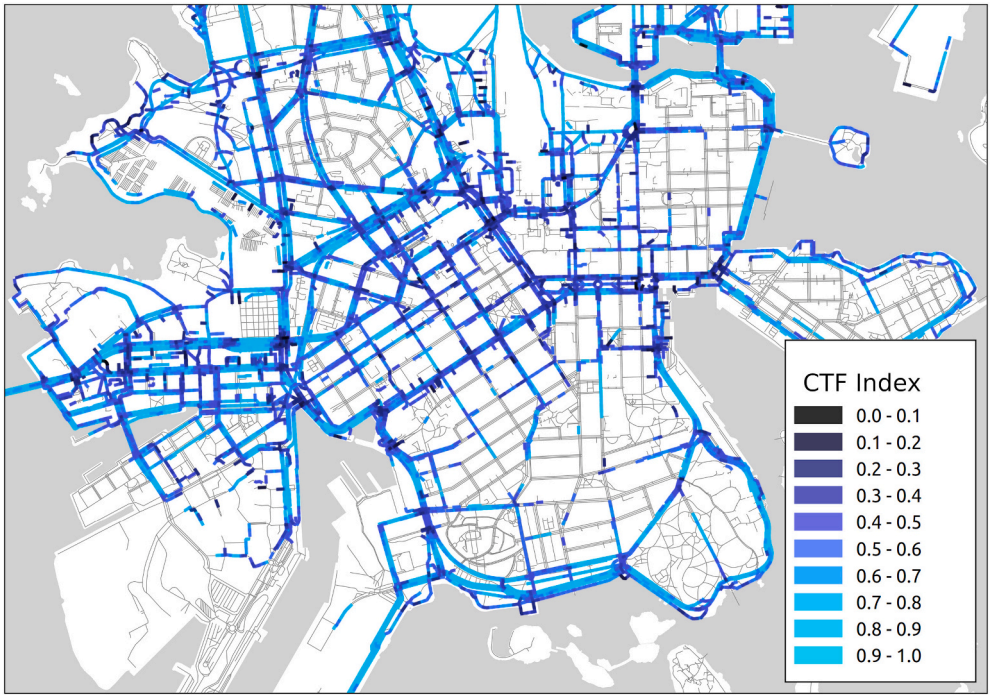


Fig. 11. Spatial distribution of the cycling traffic fluency index I_{fluency} in central Helsinki.

additional local minima occur. On some segments, they appeared due to longer-lasting stops, on others because of a lower speed ratio.

The mean index values for different types of infrastructure seem to reflect the differences in their cyclability. The variation is small, yet significant. The data suggests, for example, that a rough surface on walk- and cycleways is linked to an adverse effect on fluency. I_{speed} and I_{acc} were, on average, lower for segments on cycleways with an uneven surface than for segments on even cycleways. Moreover, on-street cycling on side roads appeared to be more fluent than on main roads, primarily because the cyclists stopped longer and more frequently on major streets. The infrastructure group with the lowest average for I_{stop} and its component indices were cycle lanes. Curiously, this group was also the group with the highest average values for I_{speed} and I_{acc}.

5.5. CTF as a routing criterion

Experiments with I_{fluency} as a routing criterion showed that in comparison with the shortest distance, shortest time, and highest popularity

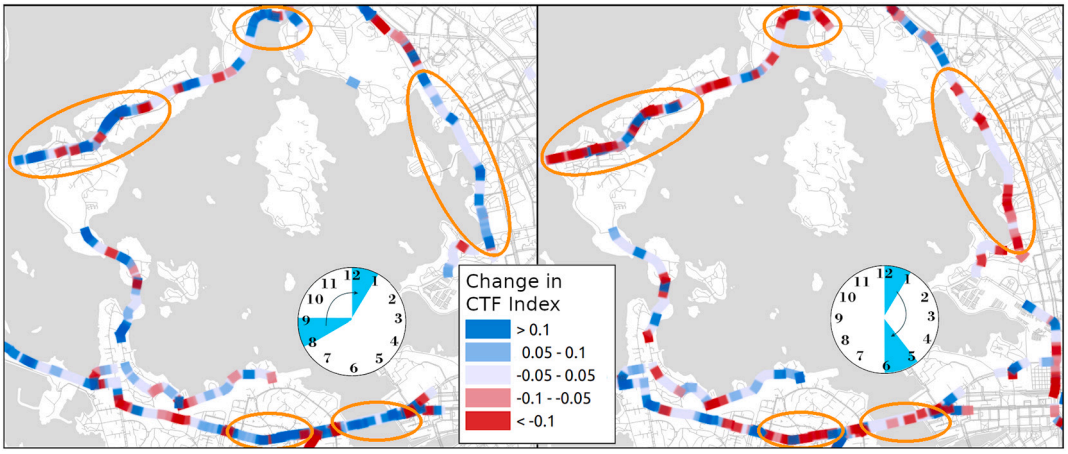


Fig. 12. Change in I_{fluency} throughout the day. From the morning rush hour to noon, I_{fluency} shows improvements for the cycling traffic towards Helsinki's city centre (left). The effect is locally restricted to certain areas that are marked with orange ellipses. This is reversed from noon to the afternoon rush hour (right).

criteria, fluency-based routing is placed on the middle ground between popularity and time (Fig. 13). Popularity-based routing sticks to very frequently traversed ways, e.g. big streets with a convenient cycling infrastructure and dedicated cycleways. In exchange for using popular infrastructure, long travelling distances are accepted. The fastest route is often fairly similar to the shortest one, but favours infrastructure that facilitates faster cycling. It is usually more continuous and has fewer sharp turns. The most fluent route tends to follow the fastest one, but accepts even more detours for more continuity, better infrastructure, and fewer turns. It hence has a surprisingly high agreement with popularity-based routing, even though I_{fluency} does not possess any notion of popularity.

6. Discussion

The method for CTF estimation presented in this article scales linearly with respect to the number of trajectories and can therefore be applied to larger trajectory volumes. It can also be adopted for similar datasets, although some processing steps may require adjustments. Depending on the GNSS sampling rate and the cleanliness of the raw trajectories, for example the degree of trajectory smoothing can be raised or reduced. For low sampling rates or very noisy trajectories, considering a more fault-tolerant method for speed and acceleration determination may be necessary.

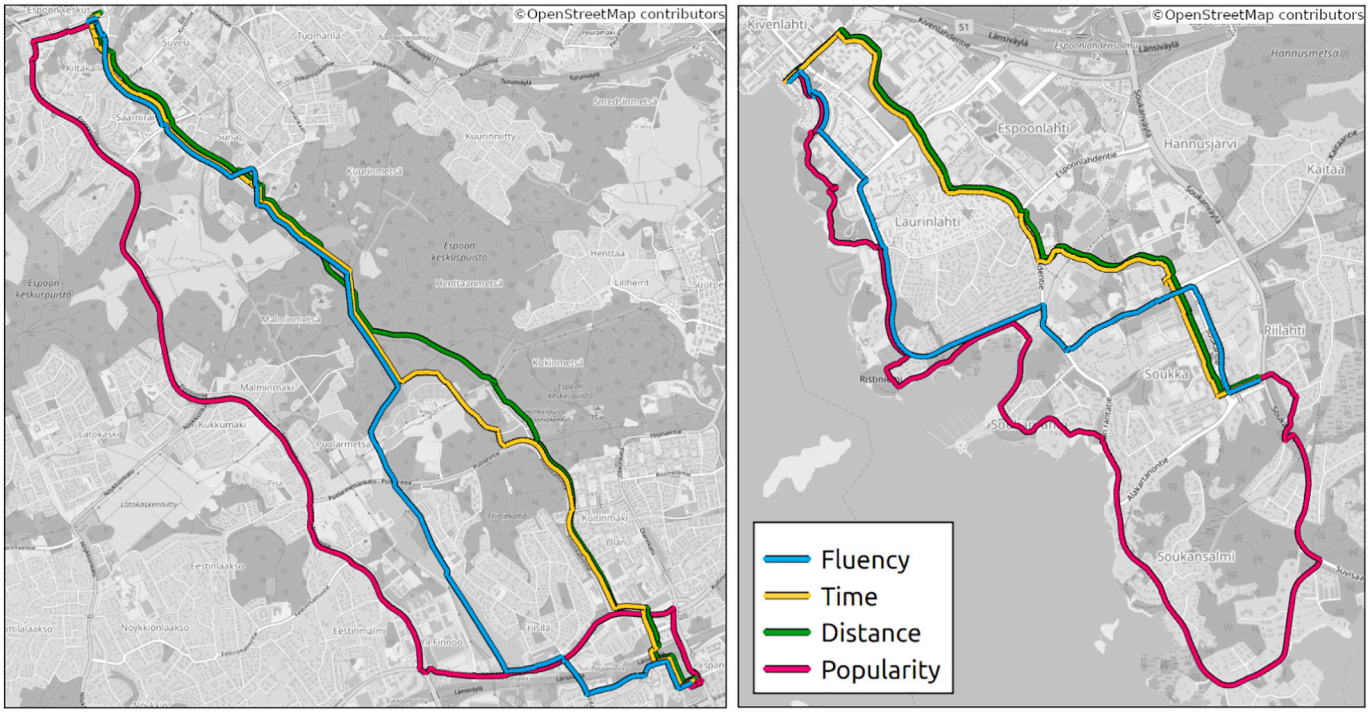


Fig. 13. Fluency-based routing (blue) is a compromise between using the shortest route (green) and infrastructure with good overall cyclability. It thus bears some resemblance to popularity-based routing (red). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

6.1. Veracity of the trajectory data

One of the key issues we faced when using big mobile tracking data to estimate CTF was the veracity of the data. We proposed a number of approaches to evaluate whether the derived information corresponded to the factual situation on the street. Positive results are a promising sign, while negative results give a general idea of potential challenges associated with big cycling trajectory data. A set of trajectories recorded in a controlled setting with demographic information and feedback from the cyclists could provide more solid proof. Our approach is, however, a much more affordable alternative.

The results of our analyses indicate a high level of veracity of the data. The continuity of segment speeds on the same infrastructure, for instance, speaks against a large influence of randomness. Further evidence stems from the high correlation between stops and intersections, especially traffic-light-regulated ones.

Nevertheless, bias and uncertainty remain a big concern. Uncertainties arise e.g. because of GNSS errors, in cities especially in urban canyons (Thiagarajan et al., 2009). Map matching can remove some of these errors (Dalumpines & Scott, 2011), but at the same time, mismatches introduce new uncertainties. On the other hand, given the volume of the dataset the impact of small errors in single trajectories is expected to be negligible, in contrast to the bias that the data carries.

Our results hint at the presence of two different kinds of bias. Considering the heavy participation inequality and the fact that some types of cyclist are more likely to use mobile tracking applications than others, we can conclude that the trajectory dataset does not represent all types of cyclist equally. Still, we find evidence that the dataset contains trajectories representing different types of cycling, most prominently in places where a street and a pavement run in parallel and both are frequently traversed. Often, the average speed of the two ways differs significantly, while the speed ratio exhibits similar values. The street-using cyclists thus seem to differ inherently from the pavement-using cyclists, which could indicate the presence of different types of cyclist.

The second type of bias manifests in the divide of the popularity of segments. A few streets and paths were disproportionally popular compared to the rest of the street network. We assume that CTF corresponds with cyclists' preferences and know that cyclists tend to favour more convenient infrastructure. Consequently, it is doubtful that unpopular streets that are obviously avoided by the majority of cyclists would be more fluent than cycling infrastructure that is known for its good cyclability. We cannot deduce whether cycling on this infrastructure is indeed more fluent and cyclists avoid it, for example for safety reasons, or if the few passing trajectories represent a type of cycling that is more fluent by nature.

Another source of uncertainty in our results are changes to the on-street circumstances. Road constructions or changes of the traffic management, for example, can change the cyclability temporarily or permanently. This is one possible explanation for stop hot spots that appear in seemingly random locations.

6.2. Suitability of the data for CTF estimation

Pertaining to the estimation of cycling traffic fluency, the question we need to ask is whether the degree of uncertainty and bias in the data is still acceptable. A paramount positive indicator is the correlation between the segment characteristics and properties of individual trajectories. Considering that the behaviour of an individual cyclist will always deviate from the average, the correlation is surprisingly high. It shows that the bias and heterogeneity of the tracks in the dataset do not invalidate its usage to estimate the behaviour of cyclists.

On the other hand, we note that not every cycling trip is equally suitable for CTF estimation, as not all cyclists are equally concerned about continuous, steady travelling. This assumption is supported by some very significant hot spots that are obviously caused by voluntary stopping behaviour. Most are located in scenic places close to the sea, e.g. in the middle of a bridge. Left untreated, behaviour that falsely indicates cycling obstructions can distort the results.

The impact of any single trajectory on a segment's characteristics is reduced the higher the trajectory count of the segment is. If the trajectory density is high, it is less crucial that each trajectory is suitable for the task, and the confidence in the derived information rises. We find that the minimum number of trajectories per segment could be raised even higher than the current value of 10, because local extrema of the CTF index that are hard to explain often coincide with a sparse trajectory density.

Then again, we observe that CTF in summer, derived from hundreds of trajectories, and winter, based on only tens of trajectories, significantly varies only in a small number of locations, in spite of the large difference in the data density. Surprisingly, the indicated CTF tends to be higher in winter due to higher segment speed values. One explanation could be that exercise-oriented, and thus more confident cyclists are more likely to ride in less optimal weather and street conditions (Bergström & Magnusson, 2003), and presumably the share of utilitarian trips is higher in the winter. Consequently, the variation of I_{fluency} would occur only partly because of a change of the circumstances on the street.

In conclusion, if the type of cyclist and the mode of cycling were known, about ten trajectories could be enough to estimate CTF. In the absence of this information, however, there seems to be no strong argument against using big mobile tracking data if it is derived from at least a few tens of trajectories.

6.3. Conception of the CTF index

Due to its modular design, the CTF index is a highly adaptable measure. Through modification of the transformation functions of the index components, the degree of fluency indicated by the input characteristics can be customized. By changing the index combination functions, the influence of the different components on the final CTF estimation can be altered. As the measure incorporates only fundamental characteristics of the cycling traffic flow, its general concept is not tailored to the study region and can be applied to any urban region.

Designed to facilitate visual analysis, the presented index emphasizes subtle differences, for example through the sharp distinction of below and above average speed values. Through the categorisation of the stop-related characteristics, it furthermore incorporates elements of simplification. For other applications, e.g. if the index is used as a routing criterion, the transformation functions can be made more robust by eliminating abrupt value changes, for example by replacing the stop-related characteristics' staircase functions with continuous functions.

The CTF index implements the preferences of cyclists suggested by preference studies. Accordingly, the chosen transformation functions favour smooth travelling and penalize interruptions and unsteady movement. Judging from the index's ability to reflect variation in the data, it provides a good starting point for the estimation of CTF. Some configuration details, however, can be subject to discussion. For example, it could be argued that the current penalization of stops with a short duration is too mild, so that their impact on the cycling quality is not properly reflected. A definite conclusion can only be reached by incorporating a notion of what cyclists themselves perceive as fluent cycling. In future work, this could be achieved by means of a survey that would investigate the perception of cycling in different real-world conditions.

7. Conclusions

This paper presents one possibility for utilizing mobile activity tracking data to characterize cycling in urban environments. For this purpose, it introduces the concept of cycling traffic fluency (CTF), i.e. the smoothness of the cycling traffic flow. In a multi-stage procedure that uses a large set of cycling trajectories as input, characteristics describing the dynamics and stopping behaviour of cyclists on segments

of the street network are derived. By combining these characteristics, a quality index is obtained that indicates a high level of CTF wherever cyclists travel continuously at a comfortably fast speed.

The components of the index, particularly its transformation and combination functions, are customizable for various applications. Due to the conformity of the concept with important preferences stated by cyclists, the index could be adapted and utilized as a multifaceted routing criterion. Furthermore, it has the potential to serve as a decision criterion for city planners, supporting the identification of single streets or whole neighbourhoods that are in need of improvement. In other words, the index could be used to convey a picture of the cyclability of a city.

Major advantages of using big mobile activity tracking data for cycling traffic analysis are the data's large spatio-temporal extent, resolution, and number of contributing cyclists. It can therefore be a powerful base for identifying large-scale spatio-temporal patterns in urban areas. However, the data is also prone to uncertainty and different kinds of bias that reduce the reliability of the CTF estimate. Significant variation in the popularity of different streets, especially in combination with the participation inequality, limits the representativeness of the data. Additionally, the comparability of CTF on different types of infrastructure is reduced by inherent differences between the associated sets of trajectories which suggest that the prevailing types of cycling differ. Besides this, the data shows patterns that are not related to CTF but deliberate choices of cyclists.

On routes where the data density is higher than a few tens of trajectories per segment, noise becomes negligible and the impact of misleading trajectory properties on the CTF estimate is contained. Especially for segments traversed by fewer trajectories, the accuracy and representativeness of the estimate could be increased by means of metadata-supported trajectory weighting. In future work, ways to infer missing metadata, e.g. the type of cyclist and the trip purpose, could be developed. Moreover, a costly yet expedient measure to validate and improve the presented method would be to reconcile the estimation with the perceptions of cyclists.

Declaration of Competing Interest

None.

Acknowledgements

We are grateful to Sports Tracking Technologies Ltd. for granting us the access to the data used in this work. Only tracks that were made public by the users were included in the dataset we received.

References

Baker, K., Ooms, K., Verstockt, S., Brackman, P., De Maeyer, P., & Van de Walle, R. (2017). Crowdsourcing a cyclist perspective on suggested recreational paths in real-world networks. *Cartography and Geographic Information Science, 44*(5), 422–435. https://doi.org/10.1080/15230406.2016.1192486.

Bergman, C., & Oksanen, J. (2016a). Conflation of OpenStreetMap and mobile sports tracking data for automatic bicycle routing. *Transactions in GIS, 20*(6), 848–868. https://doi.org/10.1111/tgis.12192.

Bergman, C., & Oksanen, J. (2016b). Estimating the biasing effect of behavioural patterns on Mobile fitness app data by density-based clustering. In L. T. S. T. Sarjakoski, & M. Y. Santos (Eds.), *Geospatial data in a changing world – Selected papers of the 19th AGILE conference on geographical information science* (pp. 199–238). Switzerland: Springer. https://doi.org/10.1007/978-3-319-33783-8 12.

Bergström, A., & Magnusson, R. (2003). Potential of transferring car trips to bicycle during winter. *Transportation Research Part A: Policy and Practice, 37*(8), 649–666. https://doi.org/10.1016/S0965-8564(03)00012-0.

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

Bundesministerium für Verkehr, Innovation und Technologie. (2017). Kosteneffiziente Maßnahmen zur Förderung des Radverkehrs in Gemeinden. https://www.bmk.gv.at/dam/jcr:acab66d6-cc76-4fe7-80a1-84dcb358a11e/radverkehrsfoerderung.pdf (Accessed 3 April 2020).

Caulfield, B., Brick, E., & McCarthy, O. T. (2012). Determining bicycle infrastructure preferences–a case study of Dublin. *Transportation Research Part D: Transport and Environment, 17*(5), 413–417. https://doi.org/10.1016/j.trd.2012.04.001.

Commonwealth of Australia. (2018). *National road safety action plan 2018–2020*. https://www.roadsafety.gov.au/sites/default/files/2019-11/national road safety action plan 2018 2020.pdf. (Accessed 3 April 2020).

Dalumpines, R., & Scott, D. M. (2011). GIS-based map-matching: Development and demonstration of a postprocessing map-matching algorithm for transportation research. In *Advancing geoinformation science for a changing world* (pp. 101–120). Springer. doi:https://doi.org/10.1007/978-3-642-19789-5 6.

Damant-Sirois, G., Grimsrud, M., & El-Geneidy, A. M. (2014). What's your type: A multidimensional cyclist typology. *Transportation, 41*(6), 1153–1169. https://doi.org/10.1007/s1111601495238.

Dane, G., Feng, T., Luub, F., & Arentze, T. (2019). Route choice decisions of e-bike users: Analysis of GPS tracking data in the Netherlands. In *The annual international conference on geographic information. science*, 109–124. https://doi.org/10.1007/978-3-030-14745-7 7.

De Mauro, A., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review, 65*(3), 122–135. https://doi.org/10.1108/LR-06-2015-0061.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik, 1*, 269–271. https://doi.org/10.1007/BF01386390.

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the second international conference on knowledge discovery and data mining (KDD-96) (pp. 226–231). AAAI Press. doi:10.1.1.121.9220.

Ferrari, L., & Mamei, M. (2013). Identifying and understanding urban sport areas using Nokia sports tracker. *Pervasive and Mobile Computing, 9*(5), 616–628. https://doi.org/10.1016/j.pmcj.2012.10.006.

Forney, G. D., Jr. (1973). The Viterbi algorithm. *Proceedings of the IEEE, 61*, 268–278. https://doi.org/10.1109/PROC.1973.9030.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137–144. https://doi.org/10.1016/j.ijinfomgt.2014.10.007.

Griffin, G., & Jiao, J. (2015). Crowdsourcing bicycle volumes: Exploring the role of volunteered geographic information and established monitoring methods. *URISA Journal, 27*(1), 57. https://doi.org/10.31235/osf.io/e3hbc.

Griffin, G., Nordback, K., Götschi, T., Stolz, E., & Kothuri, S. (2014). Monitoring bicyclist and pedestrian travel and behavior: Current research and practice. *Transportation Research Circular*. https://doi.org/10.17226/22420 (E-C183).

Harvey, F. J., & Krizek, K. J. (2007). *Commuter bicyclist behavior and facility disruption (tech. Rep. No. report no. MnDOT 2007–15)*. University of Minnesota.

Hochmair, H. H., Bardin, E., & Ahmouda, A. (2019). Estimating bicycle trip volume for Miami-Dade county from Strava tracking data. *Journal of Transport Geography, 75*, 58–69. https://doi.org/10.1016/j.jtrangeo.2019.01.013.

Hood, J., Sall, E., & Charlton, B. (2011). A GPS-based bicycle route choice model for San Francisco, California. *Transportation letters, 3*(1), 63–75. https://doi.org/10.3328/TL.2011.03.01.63-75.

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

Newson, P., & Krumm, J. (2009). 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). https://doi.org/10.1145/1653771.1653818.

Oksanen, J., Bergman, C., Sainio, J., & Westerholm, J. (2015). Methods for deriving and calibrating privacy-preserving heat maps from mobile sports tracking application data. *Journal of Transport Geography, 48*, 135–144. https://doi.org/10.1016/j.jtrangeo.2015.09.001.

Palma, A. T., Bogorny, V., Kuijpers, B., & Alvares, L. O. (2008). A clustering-based approach for discovering interesting places in trajectories. In *Proceedings of the 2008 acm symposium on applied computing* (pp. 863–868). https://doi.org/10.1145/1363686.1363886.

Primault, V., Boutet, A., Mokhtar, S. B., & Brunie, L. (2018). The long road to computational location privacy: A survey. *IEEE Communications Surveys & Tutorials, 21*(3), 2772–2793.

Pucher, J. R., & Buehler, R. (2012). *City cycling*. Vol. 11. Cambridge, MA: MIT Press.

Rao, A. M., & Rao, K. R. (2012). Measuring urban traffic congestion - a review. *International Journal for Traffic & Transport Engineering, 2*, 4. https://doi.org/10.7708/ijttÆe.2012.2(4).01.

Reddy, S., Shilton, K., Denisov, G., Cenizal, C., Estrin, D., & Srivastava, M. (2010). Biketastic: sensing and mapping for better biking. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1817–1820). https://doi.org/10.1145/1753326.1753598.

Rodgers, J. L., & Nicewander, W. A. (1988). Thirteen ways to look at the correlation coefficient. *The American Statistician, 42*(1), 59–66. https://doi.org/10.1080/00031305.1988.10475524.

Romanillos, G., Zaltz Austwick, M., Ettema, D., & De Kruijf, J. (2016). Big data and cycling. *Transport Reviews, 36*(1), 114–133. https://doi.org/10.1080/01441647.2015.1084067.

Rubin, V., & Lukoianova, T. (2013). Veracity roadmap: Is big data objective, truthful and credible? *Advances in Classification Research Online, 24*(1), 4. https://doi.org/10.7152/acro.v24i1.14671.

Schüssler, N., & Axhausen, K. W. (2008). *Identifying trips and activities and their characteristics from GPS raw data without further information* (p. 502). Arbeitsberichte Verkehrs- und Raumplanung. https://doi.org/10.3929/ethz-a-005589980.

Schüssler, N., & Axhausen, K. W. (2009). Processing raw data from global positioning systems without additional information. Transportation Research Record, 2105 (1), 28–36. doi:10.3141/2105-04.

Sener, I. N., Eluru, N., & Bhat, C. R. (2009). An analysis of bicycle route choice preferences in Texas, us. *Transportation, 36*(5), 511–539. https://doi.org/10.1007/s11116-009-9201-4.

Shen, L., & Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews, 34*(3), 316–334. https://doi.org/10.1080/01441647.2014.903530.

Smith, A. (2015). Crowdsourcing pedestrian and cyclist activity data. White paper series. http://www.pedbikeinfo.org/cms/downloads/PBIC WhitePaper Crowdsourcing.pdf. ((Accessed 3 April 2020)).

Spaccapietra, S., Parent, C., Damiani, M. L., de Macedo, J. A., Porto, F., & Vangenot, C. (2008). A conceptual view on trajectories. *Data & Knowledge Engineering, 65*(1), 126–146. https://doi.org/10.1016/j.datak.2007.10.008.

Stinson, M. A., & Bhat, C. R. (2005). A comparison of the route preferences of experienced and inexperienced bicycle commuters. In *84th annual meeting of the Transportation Research Board, Washington, DC*.

Strava, I. (2018). Strava upload rate surges 5x, total uploads surpass 2 billion. https://blog.strava.com/press/strava-upload-rate-surges-5x-total-uploads-surpass-2-billion.

Strava, I. (2019). About metro. https://metro.strava.com/faq.

Strava metro data analysis summary. (2018). https://www.codot.gov/programs/bikeped/documents/strava-analysis-summary 06–25-18.pdf. Colorado Department of Transportation. ((Accessed 3 April 2020)).

Sultan, J., Ben-Haim, G., Haunert, J.-H., & Dalyot, S. (2015). Using crowdsourced volunteered geographic information for analyzing bicycle road networks. In *International Federation of Surveyors, article of the month December 2015*.

Thiagarajan, A., Ravindranath, L., LaCurts, K., Madden, S., Balakrishnan, H., Toledo, S., & Eriksson, J. (2009). Vtrack: Accurate, energy-aware road traffic delay estimation using mobile phones. In *Proceedings of the 7th ACM conference on embedded networked sensor systems* (pp. 85–98). https://doi.org/10.1145/1644038.1644048.

Wang, Z., He, S. Y., & Leung, Y. (2018). Applying mobile phone data to travel behaviour research: A literature review. *Travel Behaviour and Society, 11*, 141–155. https://doi.org/10.1016/j.tbs.2017.02.005.