Landscape and Urban Planning 232 (2023) 104686

Contents lists available at ScienceDirect

Landscape and Urban Planning

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

Bias and precision of crowdsourced recreational activity data from Strava

Zander S. Venter [*], Vegard Gundersen, Samantha L. Scott, David N. Barton

Norwegian Institute for Nature Research – NINA, Sognsveien 68, 0855 Oslo, Norway

HIGHLIGHTS

- Strava captures spatial and temporal variation in recreational activity accurately.
- Under-representation of young, elderly, and low socioeconomic status groups.
- Trend analyses need to account for growth in Strava usership and time of year.
- Adoption of Strava in urban planning depends on precision/bias requirements.

ARTICLE INFO

Keywords:
Accuracy
Mobility
GPS tracking
Physical activity
Green space

ABSTRACT

Recreational activity is the single most valuable ecosystem service in many developed countries with a range of benefits for public health. Crowdsourced recreational activity data is increasingly being adopted in management and monitoring of urban landscapes, however inherent biases in the data make it difficult to generalize patterns to the total population. We used in-situ observations and questionnaires to quantify accuracy in Strava data - a widely used outdoor activity monitoring app – in Oslo, Norway. The precision with which Strava data captured the spatial ($R^2 = 0.9$) and temporal variation ($R^2 = 0.51$) in observed recreational activity (cyclist and pedestrian) was relatively high for monthly time series during summer, although precision degraded at weekly and daily resolutions and during winter. Despite the precision, Strava exhibits significant biases relative to the total recreationist population. Strava activities represented 2.5 % of total recreationist activity in 2016, a proportion that increased steadily to 5.7 % in 2020 due to a growing usership. Strava users are biased toward cyclists (8 % higher than observed), males (15.7 % higher) and middle-aged people (20.4 % higher for ages 35–54). Strava pedestrians that were able to complete a questionnaire survey (>19 years) were biased to higher income brackets and education levels. Future studies using Strava data need to consider these biases – particularly the under-representation of vulnerable age (children/elderly) and socio-economic (poor/uneducated) groups. The implementation of Strava data in urban planning processes will depend on accuracy requirements of the application purpose and the extent to which biases can be corrected for.

1. Introduction

Recreation is arguably the most economically valuable ecosystem service in many developed countries due to the benefits it has for mental, physical and emotional wellbeing which mitigate the public health burden (Davies & Dutton, 2021; Hermes et al., 2018). The value of recreational spaces in urban settings has been exemplified by the drastic increase in outdoor recreation witnessed during the COVID-19 pandemic (Day, 2020; Samuelsson et al., 2020; Venter et al., 2021). The ability to measure and monitor recreational activity is important for designing, planning and managing resilient and sustainable cities for the following reasons: 1) data on outdoor recreation informs socio-ecological science and natural resources impact analyses on the relationship between people and their recreational activities in a given landscape setting (Hansen, 2021); 2) research outputs on recreational activity can inform data-driven decision making and management strategies in a range of application domains including urban planning (e.g. targeting green infrastructure planting), landscape architecture (e.g. enhancing aesthetic value of a park), transport infrastructure design (e.g. walkability of roads); 3) operational recreation monitoring enables testing new management interventions in near real-time and facilitates ongoing accounting for ecosystem services, and 4) valuing recreation

* Corresponding author.
E-mail address: zander.venter@nina.no (Z.S. Venter).

https://doi.org/10.1016/j.landurbplan.2023.104686
Received 6 June 2022; Received in revised form 3 January 2023; Accepted 6 January 2023
Available online 9 January 2023

services in urban ecosystem accounting. The use of crowdsourced data for health impact assessment of changes in physical recreation activity is limited by the individual anonymization of the data.

Traditional methods to measure and monitor volume, spatiotemporal extent, type of activity and user characteristics of recreational activity and travel can be divided into manual and automated methods (Alattar et al., 2021). Manual methods include on-site observation studies, questionnaire surveys, GPS surveys, video recording, and handheld counters. Automated methods include infrared sensors, magnetometers, and pressure pads which count passers-by. Traditional data collection methods are advantageous because they can be conducted according to statistical sampling procedures that allow for design-based inference to the target population. However, they are also time consuming, costly to deploy at large scale, and survey techniques are increasingly vulnerable to steadily decreasing response rates, respondent recall errors and reporting biases (Fredman et al., 2009). The recent proliferation of social media, mobile phone apps and wearable devices that sync personal location data to the web has afforded crowdsourcing of recreational activity patterns (Bubalo et al., 2019; Byczek et al., 2018; Havinga et al., 2020). Crowdsourced data on recreation is generally inexpensive, scalable to large areas, and can yield novel insights in near real-time. However, there are challenges such as the requirement for advanced data science skills to process the data, varying levels of accessibility and data sovereignty depending on how crowdsourced tools are funded, representativeness of data and the potential impact of biased data on equity in decision making (Nelson et al., 2021; Niu & Silva, 2020).

Perhaps one of the most promising forms of crowdsourcing with respect to recreational activity is outdoor activity sharing platforms including Condoon, Geocaching, GPSies, MapMyFitness, Wikiloc and Strava (Havinga et al., 2020). Here recreationists use wearable devices or mobile phone apps to track their outdoor activity and the data is uploaded to a central platform. Strava stands out from the other platforms in that it has the largest usership with over 95 million people worldwide and a growth rate of 2 million per month (Strava, 2022). Strava uses anonymized and aggregated GPS location information from its users to quantify activity over space and time at fine spatial resolution (i.e. individual trail segments) and at the global scale (https://www.strava.com/heatmap). Users define their age and gender and input the travel mode (cycling or pedestrian) they are engaging in for each activity they record using the phone's inbuilt GPS device. Due to Strava's data sharing with selected partners and public agencies, it has been adopted in scientific studies more than other outdoor activity sharing platforms. In the scientific literature Strava has been used to explore recreational use of nature areas (Thorsen et al., 2022; Venter et al., 2020); however, the bulk of studies using Strava data have focused on cycling activity including both commuting and leisure travel because of its relevance to transportation research questions (Lee & Sener, 2020). In the context of bicycle monitoring Strava has been used for travel demand estimation, route choice analysis, infrastructure evaluation, crash exposure control and air pollution exposure assessment. There remains scope for similar applications of Strava data with pedestrian (walking, running, hiking) activity and in relation to recreation landscape design, management and monitoring.

As with other forms of crowdsourced data, Strava is a small subsample of the total population and therefore needs to be calibrated with fixed-point counter stations so that one can make population-level estimates of total activity volume (Lee & Sener, 2020; Nelson et al., 2021). Furthermore, Strava is not a representative sample of the broader population and therefore using it in urban planning or policy settings may lead to inequitable outcomes. Strava data suffers from selection bias due to the type of people who select to use the platform which is currently geared to incentivise competition and, therefore, fitness-focused users are likely over-represented (Hochmair et al., 2019). Studies in North America and Australia have found that users of smartphone apps, including Strava, that track cycling activity tended to undersample females, older adults, and lower-income populations (Blanc et al., 2016; Heesch et al., 2016). Studies focusing on Strava cycling activity have found that there is an overrepresentation of middle-age and male demographics amongst Strava users (Lee & Sener, 2020). However, to date there has been no evaluation of precision and bias in Strava data that encompasses both cycling and pedestrian recreational activity. This is particularly pertinent given the growth in Strava usership and the fact that their data is now being made available to public agencies and research institutes through a dedicated data sharing platform called Strava Metro.

Due to the unknown uncertainty inherent in crowdsourced recreational data like Strava, and the need to quantify the uncertainty for robust data-driven decision making in urban planning, here we aim to quantify the precision and bias of Strava recreation activity data for those who visit nature areas in Oslo, Norway. Given that previous work on Strava data biases has focused on cycling activity for travel purposes (both leisure and commuting), our unique contribution is to quantify biases for both pedestrian and cycling activity while focusing on non-commute, recreational trips. In the context of outdoor recreational activity, the inclusion of pedestrian activities is important given that it captures a wide range of non-cycling activity (e.g. hiking, dog walking) which are important to account for in urban recreation area planning and management. For the purposes of this analysis, we define precision as a measure of how similar the spatial and temporal variation in activity estimates (i.e. Strava) are to those of the true activities. We define bias as a measure of how different any one estimated activity count is from the true activity count (also called systematic error). Both bias and precision are components of accuracy. We use fixed-point counter stations (active between 2016 and 2020), and observation and questionnaire surveys (collected during 2021) as reference data to compare Strava against. Specifically, we aim to quantify (1) how precisely Strava captures temporal and spatial patterns in recreational activity; (2) how biased the absolute number of Strava activities is and how this changes year-on-year; (3) how biased the Strava data is across activity types, gender, age, wealth and education status.

2. Methods

2.1. Study area

The study was conducted in Oslo and its surroundings (Fig. 1). Oslo is Norway's capital city (59′55 N, 10′45 E) with a population of 699 827 which accounts for approximately 13 % of the country's population (Statistikkbanken, 2022). The built-up zone of the city is punctuated with green spaces and riverine corridors which extend outward to an intact forest zone surrounding the city called "Marka" (altogether 1700 km^2) which is protected from urban development by law. The regular Marka recreation survey of Oslo households records 26 recreation activities (Kantar, 2021) – in 2020 85 % of households reported walking in Marka peri-urban forest during the year, 54 % cross-country skiing, 42 % jogging/running, and 25 % reported biking. The Oslo region provides a range of recreation opportunities that exist along a steep gradient from urban to wilderness areas. Due to the unique access to surrounding semi-natural forest and nature reserves, Oslo is not representative of many European cities, but is conceptually representative of a recreational gradient and diversity of recreation area types that is accessible to a diversified urban population.

2.2. Strava mobility data

Strava make their data available through their Strava Metro Service to selected partners. The data service is a web-based interactive platform to explore, query and download activity data. In order to get access to the Strava Metro platform, partners are required to enter into a license agreement which generally restricts the use of the data to use-cases which inform the planning or maintenance of transportation

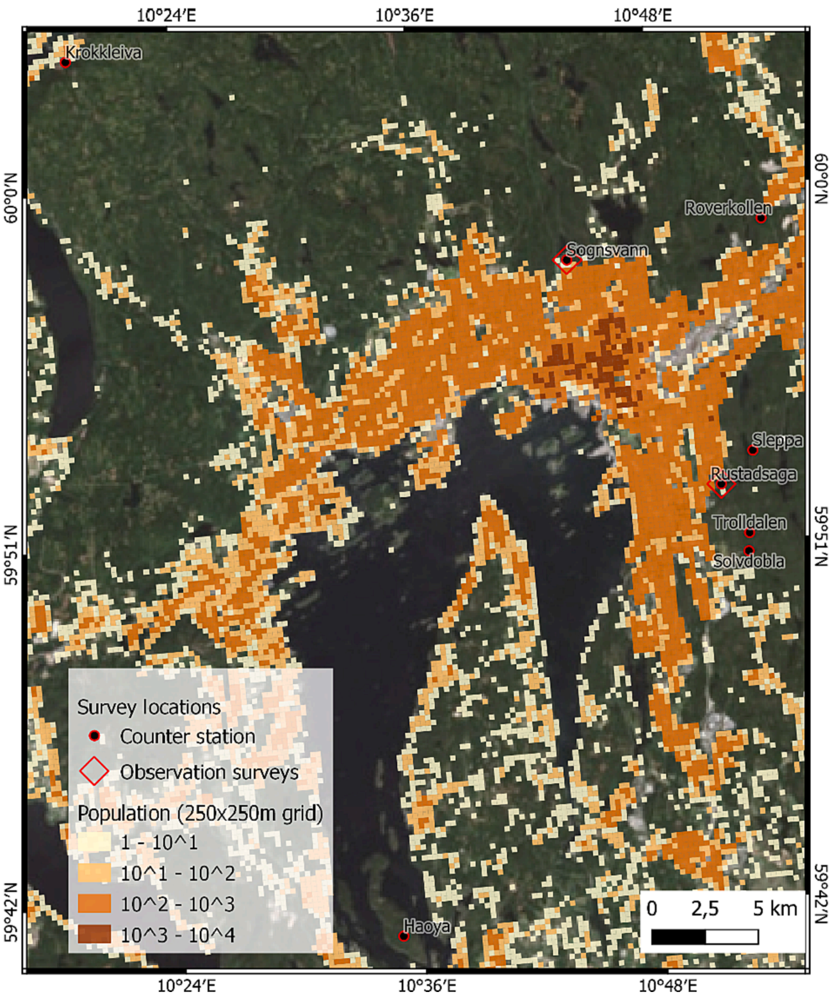

Fig. 1. Study area in Norway including the distribution of fixed-point counter stations and observation surveys. Population density data from Statistics Norway (Statistikkbanken, 2022) is mapped for reference and a satellite image base layer is used to differentiate water, forest and open vegetation.

infrastructure and processes for bicyclists and pedestrians. Selected partners include public agencies such as Departments of Transportation, Metropolitan Planning Organizations, and in some cases research institutions. To maintain user anonymity and privacy, the data is aggregated and de-identified, consistent with the European Union's GDPR and the California Consumer Privacy Act (CCPA).

The mechanism behind the Strava data is the use of a mobile phone's in-build GPS device to record the location of the phone over time and space. The raw GPS tracks from individual mobile phones are snapped to the closest recreational or transport line geometry defined by an OpenStreetMap (https://www.openstreetmap.org) base layer. Activities per line geometry are aggregated to hourly, daily, monthly and yearly time intervals if there were at least three unique users (a data privacy measure) during the given time window. Activity counts are stratified by the gender, the age (in brackets of 13–19, 20–34, 35–54, 55–64, and greater than 65), the type of activity (pedestrian or cycling) and the purpose of the activity (commuting for work or leisure). These stratifications are based on user-defined metadata collected through the Strava app. We restricted our analysis to leisure trips because the focus of our study was recreational use. All daily Strava activity counts between 2016 and 2020 for the trail segments intersecting counter station locations (Fig. 1) were downloaded from the Strava Metro Service. In addition we downloaded activity counts, stratified by user type and demographic, for the days in June 2021 and trail segments which coincided with our observational surveys (see section 2.4).

2.3. Counter stations

To quantify the temporal and spatial accuracy of Strava data we collected reference data on recreational activity numbers along selected trail segments where fixed-point counter stations had been installed (Fig. 1).The counter stations (EcoCounter with two-way pyroelectric sensor) are installed and managed for high counter accuracy following standard procedures for installation and management (Andersen et al., 2014) and record the number of people passing by per hour. Counter error rates under normal conditions have been tested to be accurate within 5 %; however, error rates increase with increasing recreational traffic.

Data from two counter stations managed by Oslo municipality at two of the most popular trail heads in Oslo, namely Sognsvann and Rustadsaga were collocated with our observation and questionnaire surveys (Fig. 1). To quantify the accuracy of the spatial variation in Strava activity counties, we required additional reference data spread over space. Therefore, we included data from five additional counter stations maintained by Statens naturoppsyn (SNO) which is a governmental authority responsible for monitoring of nature protection areas. These stations are spread over the broader Oslo region and represent a gradient of use-intensity ranging from 5 to 3368 activities per day (Fig. S1). All counter stations collected daily activity counts 365 days a year and were permanently located at the positions outlined in Fig. 1. We extracted the daily counts for the time period between 2016 and 2020 which coincided with the availability of the Strava data.

2.4. Observation surveys

To explore the activity type, gender and age biases in the Strava data, we performed in-situ systematic moment observations of recreationists at high recreational traffic locations in Oslo including Sognsvann and Rustadsaga (Fig. 1; Fig. S1). At each location we randomly selected three trail segments that contained enough Strava activity to ensure data availability. We defined these as any trail with average Strava activity greater than the 75th percentile for the trails within a 3 km radius of that location. We chose high activity trails to ensure that we could collect enough observational data given the time limitations for fieldwork in the budget for our research project. Without sampling high activity trails (>75th percentile), we would not have enough data points to quantify age and gender biases in the Strava data. Observational surveys took place during daylight hours over the course of three weeks during June 2021. Survey events were randomly distributed over locations, days of the week, and hours of day in order to reduce potential bias introduced by anomalous weather. We counted all humans passing by and recorded their gender, age, and whether they were on a bicycle or not. We included a category of "other" for recreationists that were visually obscured or where it was difficult to distinguish age or gender characteristics. To calibrate our visual estimation of age and gender, we performed an initial sampling as a group of researchers with experience in social science observational surveys until we reached consensus on what constitutes a given age or gender group.

2.5. Questionnaire surveys

To aid in the capture of socio-demographic information that cannot be observed passively through observation surveys, we deployed anonymous on-site questionnaire surveys at the six observation surveys locations. Posters with a QR code linking to an online questionnaire survey were posted on visible notification boards and remained in place from June to December 2021. When recreationists decided to complete the survey by scanning the QR code with their mobile phones, they were presented with questions about: (1) what type of GPS activity monitoring apps they use to track their recreational activity; (2) how often they specifically use Strava; (3) whether they were on foot or bicycle; (4) their annual income bracket; and (5) their highest education level.

2.6. Statistical analysis

To quantify the temporal and spatial precision of the Strava data we regressed observed (i.e. counter station) on Strava activity counts and calculated the linear regression R^2 values. We did this for temporally and spatially aggregated activity counts, during summer (April to September) and winter (remaining months). We also created separate regressions for different levels of temporal aggregation including daily, weekly and monthly time series. We calculated Strava activity count bias for any given spatial or temporal unit of aggregation as the percentage of observed recreational activities (A_0) constituted of Strava activities (A_S).

$$Countbias = \frac{A_S}{A_o} \times 100$$

We calculated the annual trend in Strava activity count bias as the slope of a linear trend fitted to the monthly bias estimated across all counter station locations. Given we were not testing any specific hypotheses we did not specify any statistical models to determine significant effects or differences.

3. Results

3.1. Precision in temporal and spatial variation

Between 2016 and 2020 the seven available counter stations (Fig. 1) recorded 6.5 million activities, while the corresponding Strava trail segments recorded 0.17 million activities. The Strava data captured the temporal variation in the activity counts ($R^2 = 0.51$; Fig. 2A) with less precision than it captured the spatial variation ($R^2 = 0.9$; Fig. 2B). The temporal variation in activity during winter months was less correlated to Strava than during summer months (Fig. 2A), however this was not the case for the spatial variation (Fig. 2B). The precision of Strava data decreases as one increases the granularity of temporal aggregation from months to weeks to days (Fig. 3). In general, Strava appears to overestimate activities at the high end of the observed activity spectrum, and underestimate at the low-end (Figs. 2 and 3).

3.2. Bias in activity count and trend

Strava activities represented 2.5 % of total recreationist activity in 2016 and 5.7 % in 2020 reflecting an increase in representativity of 1.24 % (±0.7 % standard error) per year (Fig. 4A and B). This increasing trend differs from the relatively stable trend observed in counter station data (Fig. 4A) because it reflects the increase in Strava usership over time. As more people adopt Strava, the proportion of total recreationists that Strava represents increases over time. However, averaged over the study period, Strava activities represented 3.9 % (±1.1 %) of the recreation activities observed by counter stations. There was spatial variation in the temporal bias; specifically remote locations with lower recreational use intensity (e.g. Sleppa; Fig. 1) generally have greater proportions of Strava users and show larger increases in Strava representativity over time (Table S1). There was intra-annual variation in the Strava activity count bias with summer months showing higher representativity than winter months (Fig. 4C). Similarly, there is variation in Strava bias over the days of the week where Strava is less representative on weekends compared to weekdays.

3.3. Bias in activity type, sex, age, income and education

The observation surveys included 54 sampling hours spread randomly across the selected trail segments and captured 4184 recreational users. In comparison to observed recreationists, Strava recreationists were biased toward cyclists (Fig. 5A), males (Fig. 5B) and middle age groups (Fig. 5C). The bias in age demographics was most substantial for children/teenagers with Strava data including 23.3 % less activities than observed in this age bracket. The direction of the age and sex biases was the same for both pedestrian and cycling activities (Fig. 6), although Strava cyclists showed a larger bias toward the middle as opposed to younger age brackets compared to Strava pedestrians.

We received 1475 responses to the questionnaire survey, however the sample was heavily biased towards pedestrians (93 % survey respondents were pedestrians) and older age groups (97 % were older than 19). Therefore, we restricted the analysis of wealth and income biases in Strava data to pedestrian activities (i.e. excluding cyclists) from users above 19 years old. Of these respondents, 13 % reported having used Strava before, however a number of other GPS tracking apps/devices were also adopted by respondents (Fig. S2). We found that Strava users were biased toward high-income brackets and education levels (Fig. 7).

4. Discussion

4.1. Strava precision and bias in the broader context

The Strava users in Oslo represented 5.7 % of the total recreationist population in 2020 and this value has been steadily increasing year-on-year since 2016 probably due to the increasing Strava usership. Without any other studies quantifying this inter-annual trend we cannot compare our findings to other countries. Although Strava reports a growth rate of two million users per month (Strava, 2022), this is not broken down by country and it is therefore difficult to assess whether the growth trends in our data are representative of other countries.

We found that Strava activity patterns are relatively precise because

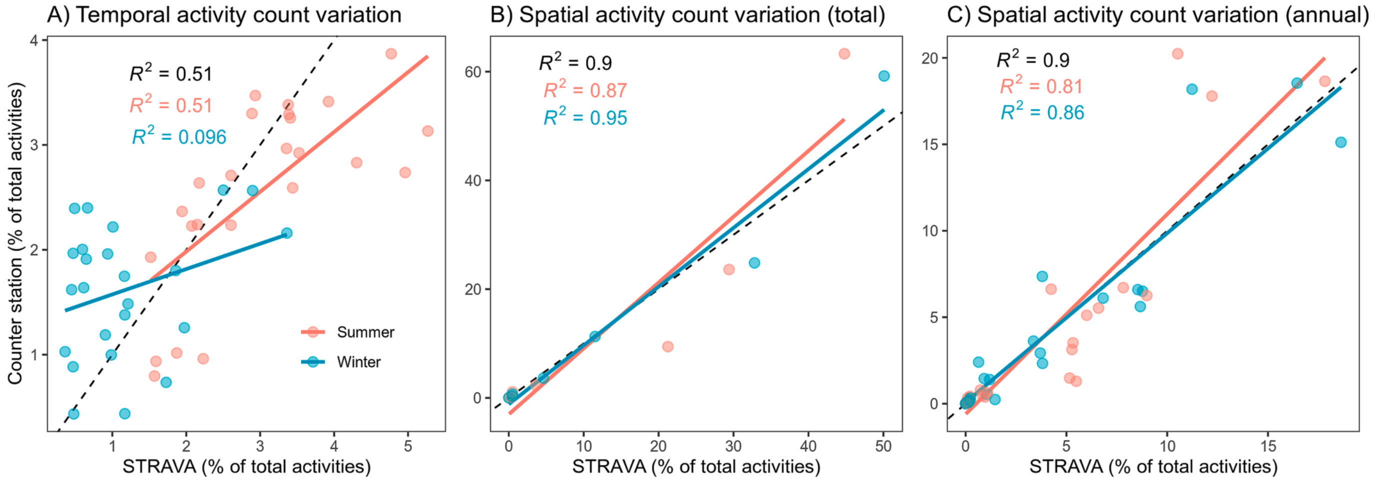

Fig. 2. Temporal and spatial correlation between Strava and counter station monthly activity counts between 2016 and 2020. Temporal variation (A) is based on counts aggregated to monthly values (n = 48), while spatial variation (B) is based on counts aggregated to unique counter locations (n = 7). C shows counts aggregated at annual interval for each unique counter location (n = 28). Values are relativized by calculating the percentage of total activity counts per grouping variable. Linear regression lines are plotted in color and a black dashed 1:1 line is added for reference. The adjusted R^2 values of linear regressions are displayed for the entire time series (in black) and separately for summer and winter.

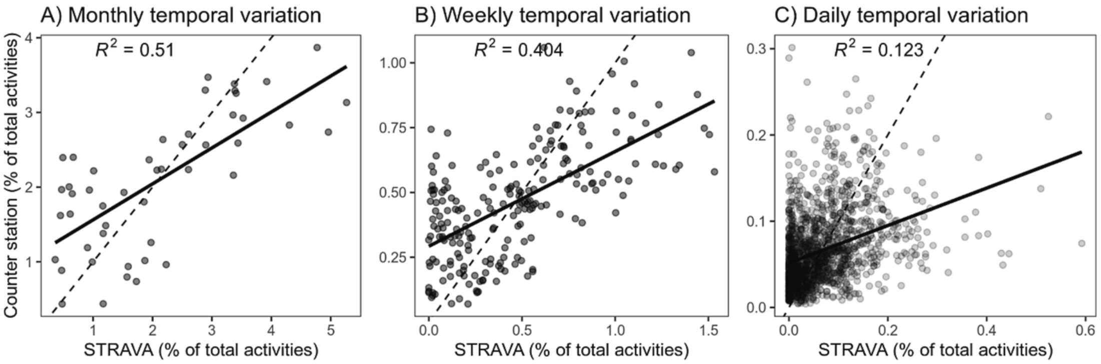

Fig. 3. Temporal correlation between Strava and counter station activity counts between 2016 and 2020 for three levels of temporal aggregation including monthly (A; n = 48), weekly (B; n = 217) and daily (C; n = 1519) sums. Values are relativized by calculating the percentage of total activity counts per grouping variable. Linear regression lines are plotted in bold along with the adjusted R^2 values. A dashed 1:1 line is added for reference.

activity counts were well-correlated with counter station data over space (R^2 = 0.9) and time (R^2 = 0.51; Fig. 2). This aligns well with the R^2 values (generally greater than 0.75) reported by a number of Strava studies reviewed in (Lee & Sener, 2020). However, none of these studies have differentiated temporal and spatial correlations, nor have they included pedestrian activity in addition to cycling activity in their analysis. We also found that temporal precision of Strava data is greater in summer than in winter and that Strava users represent a greater proportion of total recreationists during the week compared to the weekend. A potential explanation for the former pattern is that Strava users are more representative of the total recreationist population during summer compared to winter (Fig. 4C). This is possibly because during winter there are popular winter recreation activities, in particular cross-country skiing, which are not carried out by the same users as the summer pedestrian segment ofthe Strava dataset and therefore reduces correlations. The lower representativity of Strava users during the weekends may be because of the focus the Strava usership has on physically active, fitness activities which may be typically integrated with weekday routines, whereas on the weekends there are a greater variety of recreationists (e.g. foragers, campers) that are not picked up in the Strava data. A similar explanation is possible for our finding that remote locations showed less bias (greater proportion of Strava users) compared to busier, urban locations (Fig. 1). Strava users are possibly fitter than the average recreationist and are therefore able to venture further into more remote recreational areas.

We also found a reduction in Strava precision with increasing resolution of temporal aggregation (Fig. 3). This is possibly because at higher temporal units of aggregation the Strava pre-processing algorithm drops many more data points because of the privacy setting which requires at least three activities per unit time to be stored in their database. Therefore, monthly or annual aggregation would allow for a greater number of trail segments to pass the privacy threshold for data exclusion.

Our assessment of demographic and socioeconomic biases in the Strava data generally aligned with previous studies even though previous studies focused exclusively on cycling activity. We found that the direction of age and gender biases is consistent across both cycling and pedestrian activity types (Fig. 6), although Strava cyclists are more skewed to the middle-age range (between 35 and 55) compared to pedestrians which are more skewed to younger-age range (20 to 35). Lee & Sener, (2020) reviewed Strava cycling studies and reported that in almost every case Strava users are skewed towards the young-middle aged (between 25 and 44) male demographic. The authors attribute this skew to the selection bias toward economically active, tech savvy people which is a well-known bias in many other forms of crowdsourced data collected from mobile phones (Milne & Watling, 2019; Wang et al., 2018). In addition, Strava policy restricts app users to people over 13 years of age and therefore excludes the entire child demographic. The

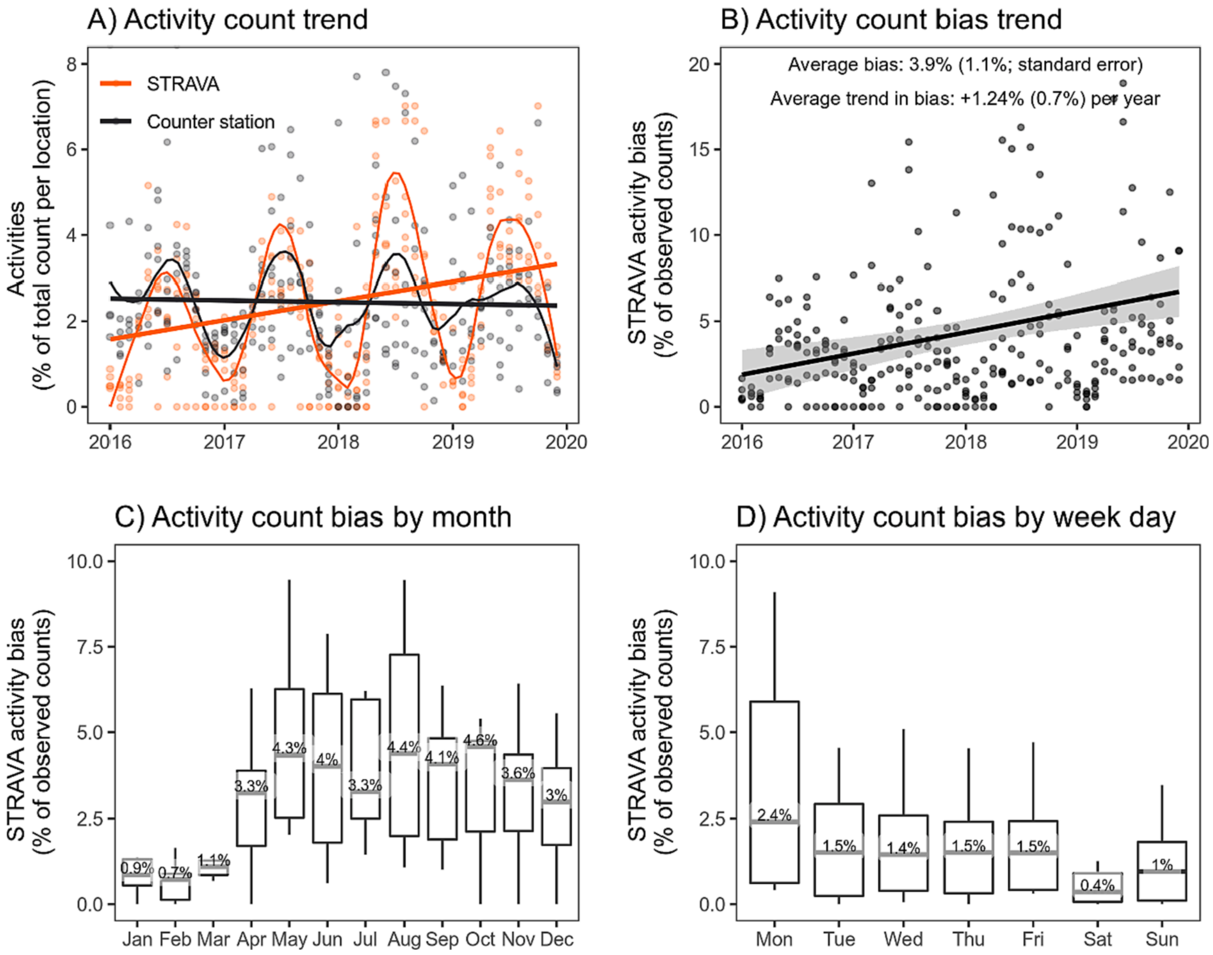

Fig. 4. Time series of relativised monthly activity counts for Strava and counter stations (A) between 2016 and 2020. A 3-month moving average line is plotted along with a linear trend line. The bias in monthly activity counts is plotted in B with a linear trend line. The activity count bias is plotted in C and D with box and whisker plots for each month and weekday, respectively. Median values are overlaid.

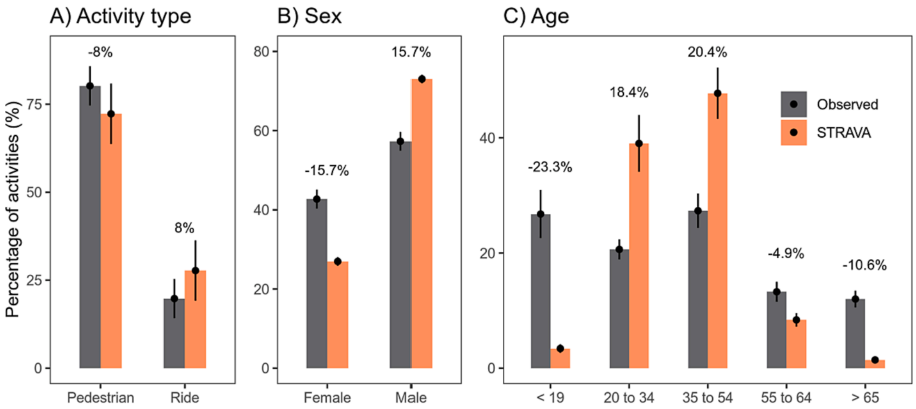

Fig. 5. Percentage composition of Strava and observed recreational users across activity type, sex and age categories. Average values are presented as bars and points with standard error bars (n = 6). The difference between Strava and observed percentages are shown above each set of bars.

same selection biases may explain why Blanc et al., (2016) found that smartphone users who tracked their cycling activity were biased toward high-income groups. In our study we found that pedestrians older than 19 years that use Strava were more likely to have higher incomes and are more educated (Fig. 7).

Apart from the inclusion of both pedestrian and cycling behavior in our bias analysis and the focus on recreational activities, our study contributes two novel aspects that have been overlooked in previous studies. Firstly, we find that Strava has a slight over-representation (8 %) of cyclists relative to pedestrians when comparing to the total recreationist population. Secondly, we find that, although the elderly are indeed underrepresented in Strava data, the largest bias is amongst teenagers and children (<19 years). The proportion of Strava users for this demographic are 4 % whereas in reality 27 % of recreationists are under 19 years of age. This is a bias also common to in-situ surveys and population surveys interviewing adults (Fredman et al., 2009).

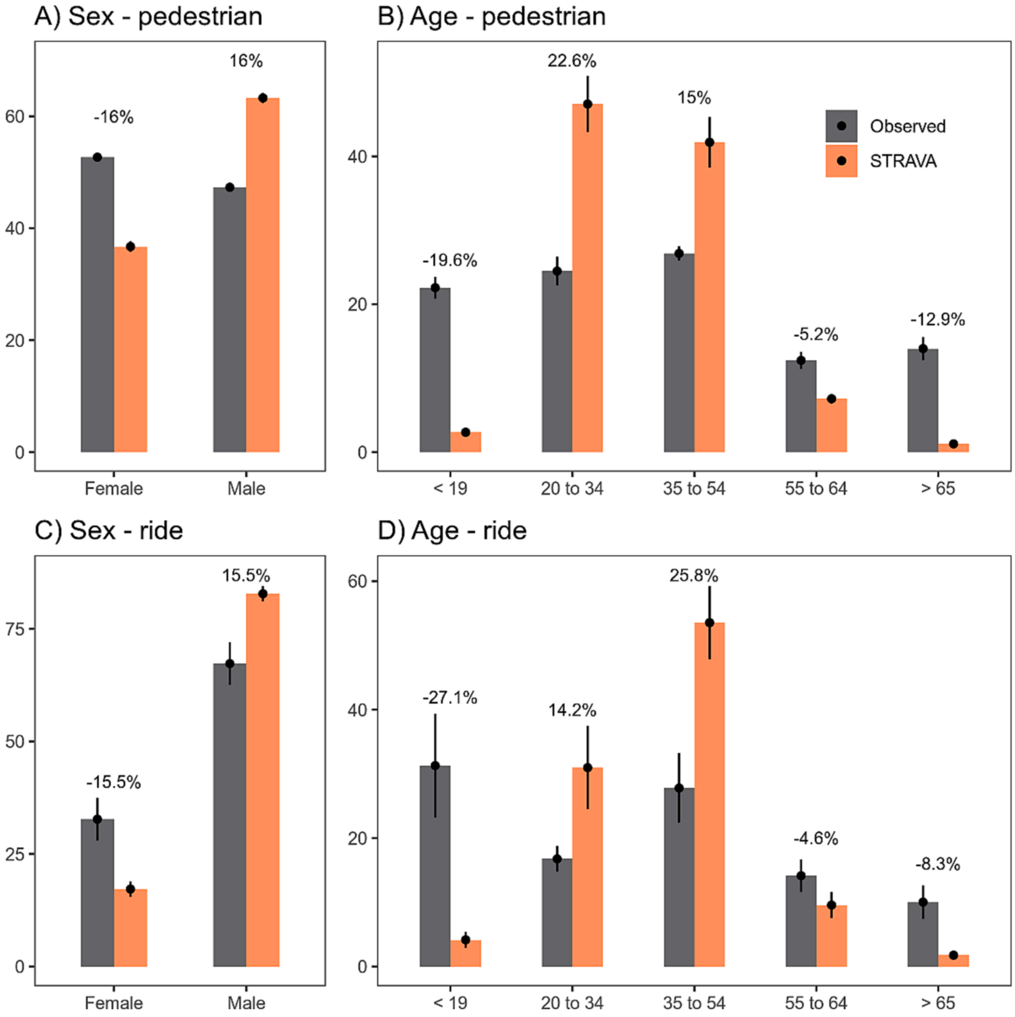

Fig. 6. Percentage composition of Strava and observed recreational users across sex (A, C) and age (B, D) categories stratified by activity type. Average values are presented as bars and points with standard error bars (n = 6). The difference between Strava and observed percentages are shown above each set of bars.

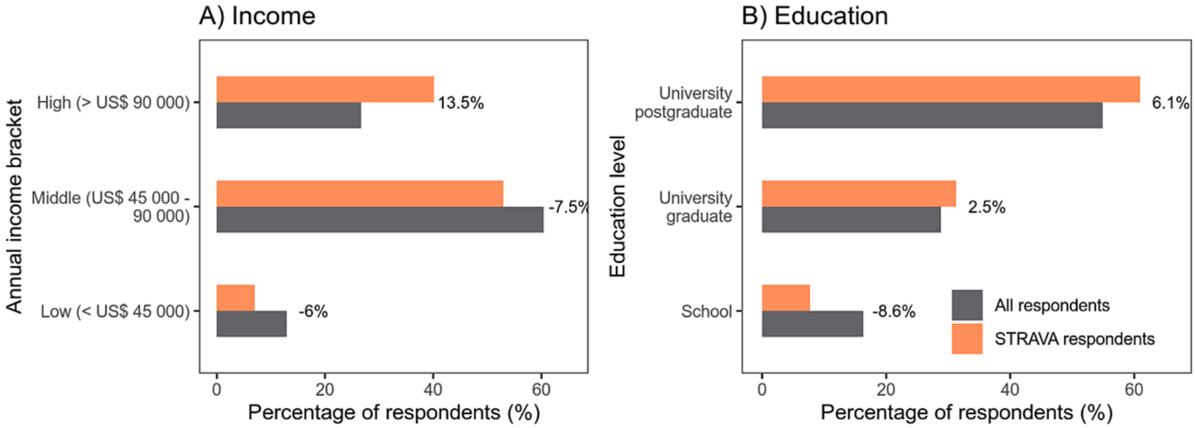

Fig. 7. Percentage composition of questionnaire respondents (pedestrians only) based on income (A; n = 889) and education (B; n = 1000) level. Compositions are calculated for respondents who use Strava and the total recreationist population (i.e. all respondents). The percentage difference between Strava and total recreationist population are shown above each set of bars.

4.2. Recommendations for use of Strava data

Any generalization about the utility of Strava data given its precision and bias is made difficult by the vast range of potential use-cases. Therefore, we suggest that reflections on how good the Strava data are be made relative to purpose requirements. For instance, using Strava within an ecosystem accounting framework to account for relative differences in recreation use intensity of green spaces may be warranted given the high spatial precision of Strava data. In this case the demographic biases are not important given inference is being made at the level of the whole population. Similarly, assessing the relative temporal changes in recreation use of different greenspaces in response to external drivers such as the pandemic lock-down measures has provided useful findings regarding social distancing and greenspace preferences in the physically active population in Oslo (Venter et al., 2020, 2021). However, if a sociologist wanted to understand route choice between different age groups using Strava data, they might find the socioeconomic and demographic biases a significant stumbling block. Nevertheless, there are a few generalizable recommendations we draw from our findings which are outlined below.

Given that our results on the precision of Strava data concur with multiple other studies, we suggest that the most robust use-case for Strava is comparing recreation use intensity between different areas and over different times. For instance, Strava activity volumes have been compared over space to explore route choice preferences and infer characteristics of urban form (e.g. park size, shape, amenities) that are associated with route selection (Alattar et al., 2021; Sun et al., 2017). Further, a study in Queensland, Australia found that Strava can accurately detect changes in cycling behavior over the short-term (3 months), and is thus useful for evaluating the effects of infrastructure change or any other sudden impact (e.g. mobility restriction) on recreationist behavior (Heesch et al., 2016). An important caveat with using Strava for trend analysis over longer time frames (i.e. years) is that one needs to account for the year-on-year increase in Strava usership (Fig. 4) so that trends in actual recreation activity are not confounded. Furthermore, users should be aware that seasonal and weekday vs weekend differences in Strava representativity should be taken into account, depending on the local context and use-case considered. For instance, researchers monitoring changes in recreational use in response to municipal interventions (e.g. installing new park amenities) should be aware that drawing conclusions about winter activity is less precise than summer activity. However, these differences may vary with local climate and population, and one might find different seasonal variation in, for instance, Mediterranean climates where winter sports like skiing are not present and may not lead to the results we found in Oslo, Norway. Therefore, it is also important to note here that any application of Strava data would do well to calibrate it with local fixed-point counter stations instead of relying on published correlations such as those reported in our study.

The socioeconomic biases in the Strava data are important to consider in applications where underrepresented groups are the focus. This is true for both cycling and pedestrian activities given the age and gender biases are present in both sets of Strava users (Fig. 6). Perhaps the most relevant application domain in this regard is epidemiology where at-risk population groups including children, elderly and low socioeconomic status groups are underrepresented in the Strava data. For example, studies that have used Strava to examine exposure to traffic related air pollution (Lee & Sener, 2019; Sun & Mobasheri, 2017) are assuming that mobility of at-risk groups matches the spatio-temporal pattern present in the Strava data. Due to the bias in the Strava data, they may be overlooking areas in the city or times of year when at-risk groups are disproportionately exposed to air pollution. Strava-data is also anonymized, making it impossible to correlate with individual level epidemiological covariates. Although cross-calibration with counter stations reveals that Strava is representative of broad-scale spatio-temporal mobility (Fig. 2), this may not be true at very local scales. For instance if counter stations were to be placed near retirement villages or kindergartens one might find the correlation with Strava activity counts breaks down significantly due to the underrepresentation of elderly and children. Similarly, due to the selection bias inherent in Strava users, any analysis of route choice or recreation area preference should be aware that the bulk of users will likely be motivated by fitness and competition outcomes instead of, for instance, aesthetic value (Dolan et al., 2021).

Apart from being cognisant of biases in the Strava data when interpreting results from analyses, it may be possible to correct for the biases by complementing Strava with other datasets. We have already discussed fixed-point counter stations at length, but socioeconomic biases may be compared with statistical population survey data or census block data. For example, (Roy et al., 2019) used statistical regression modeling with a range of spatially-explicit covariates including census data, transport network and urban form characteristics to correct for the biases in Strava data in Maricopa County, Arizona, USA. However, this effort required a significant number of counter stations (n = 104) spread across the city which may not be practical or possible in other settings. Short of correcting the bias in Strava data one can cross-reference it with statistical survey data, for example in Norway, the Marka-survey which is a representative population survey of recreational use of peri-urban forest in Oslo (Kantar, 2021). Similarly, other forms of participatory GIS can complement Strava data and fill in the gaps introduced by the socioeconomic biases. For example one could use data from the Barnetråkk (https://www.barnetrakk.no/en/) programme in Norway which allows children to map out areas of recreational interest around their kindergarten and residence.

4.3. Limitations and further research

Our results need to be interpreted in light of the scope and limitations of our study. Firstly, we focused on Strava as a source of crowdsourced recreation data due to its growing popularity and the availability of the data through the Strava Metro platform. However, as we found in our questionnaire survey (Fig. S2), there are many other sources of crowdsourced mobility data which may be used and may have their own unique biases which need to be researched. These include GPS tracking devices or apps including Garmin, Apple Health, Fitbit, Google Health, Runkeeper and more, although the availability of data from these apps is unknown at present. Additional data resources also include outdoor activity-sharing platforms which are often community-led and open-access such as Condoon, Geocaching, GPSies, MapMyFitness and Wikiloc (Havinga et al., 2020; Norman et al., 2019).

Secondly, our analysis was focused on the green space recreation at the end of the outdoor mobility spectrum and therefore does not cover more utilitarian forms of mobility (e.g. commuting to work or shops). This is reflected in our exclusion of Strava trips marked as “commute” and by the fact that we use counter stations located along trails in *peri*-urban areas demarcated for recreation (Fig. 1). Therefore, our results should not be generalized to commute Strava data or to recreational activity measured in inner-city environments. One might find, for instance, that the bias and imprecision in Strava data might be exaggerated in intra-urban parks where the range of recreational activity types and users is very different to the fitness-based Strava activity in Oslo (Evensen et al., 2021; Nordh & Østby, 2013). In this way, our assessment of spatial precision is limited to a sample size of seven and may not be robust nor representative of a broader spatial gradient which includes inner-city environments. Future work, such as studies in the literature on cycling activity (Lee & Sener, 2020), would do well to include a greater spatial distribution of reference counter stations so that we can explore how Strava biases vary with urban form and land-use. Similarly, it would be beneficial to sample Strava trails with both very low and high activity counts to explore whether age and gender biases vary with trail use intensity. In our study we selected high activity trails near recreational trail heads in order to generate enough data points for analysis given project time constraints.

Thirdly, the scope of our analysis excludes water use and off-trail recreational use due to the fact that Strava uses OpenStreetMap paths to snap to GPS positions and therefore we were restricted to using reference counter stations that are situated along established recreation paths. Excluding off-trail use and water use means that we were not able to quantify the precision and bias for certain user groups like orienteering, hunting, fishing and boating which do not necessarily utilize established paths. However, we targeted our observation and questionnaire surveys at trail heads with the aim of capturing such users because most recreationists start from a parking lot and then embark on off-trail or water trips. Yet it is likely that many off-trail users were not represented in our data. For example, 43 % of Oslo's households report swimming in lakes, 35 % report picking mushrooms and berries, 28 % camp in tents or hammocks, and 10 % do cross-country orienteering, in large part taking place off OSM trails (Kantar 2021). This is possibly reflected in the fact that Strava data was less representative of the recreationist population on weekends – when most non-exercise activities occur – compared to weekdays. However, all these activities often require walking or cycling along designated paths and therefore may be picked up by Strava activities recorded at trail heads.

Fourthly, we performed observational surveys of recreationists which necessitated visual estimation of recreationist age and gender. We attempted to mitigate sampler bias though consensus estimation within out group of researchers, and by using age categories that were relatively broad. Nevertheless, future studies would benefit from more precise measurement of demographic information through paired observation and interview survey approaches.

Finally, although we were able to differentiate age and gender biases for cycling and pedestrian activity separately, we were not able to distinguish activity type when quantifying spatial and temporal precision. This was because the counter stations used could not differentiate cyclists from pedestrians. Therefore the results for spatio-temporal precision should be interpreted as a combined (pedestrian plus cycling) estimate and future studies that are able to differentiate activity types might find differences in precision between cycling and pedestrian Strava activities.

5. Conclusion

Recognising the importance of measuring recreational activity for urban planning purposes and the problems inherent to crowdsourced data, we aimed to quantify the accuracy of Strava data in Oslo, Norway. We found that Strava can be generalizable to the entire population when quantifying relative changes over time and space, however the level of temporal aggregation (day vs month), length of study, and season of the year have important implications on precision. Strava activities underrepresented vulnerable population groups (children, elderly and low socioeconomic status groups) and therefore any application of Strava data should be cognizant of this to ensure analysis of equity is not biased. We suggest that using Strava in combination with other data sources (e.g. counter stations, population and in-situ surveys) have great potential to expand the scope and robustness of its application in recreational landscapes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The authors do not have permission to share data.

Acknowledgements

This research was carried out in support by the Research Council of Norway with project MAREA (MARine Ecosystem Accounting for integrated coastal planning in the Oslofjord; contract #326976), project LandValUse (Integrated welfare assessment of climate and biodiversity impacts of land use; contract #319917), and the project Assessing the potential for multifunctional blue-green infrastructure in southern Scandinavia (contract #281092).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.landurbplan.2023.104686.

References

Alattar, M. A., Cottrill, C., & Beecroft, M. (2021). Modelling cyclists' route choice using Strava and OSMnx: A case study of the City of Glasgow. *Transportation Research Interdisciplinary Perspectives, 9*, Article 100301. https://doi.org/10.1016/j.trip.2021.100301

Andersen, O., Gundersen, V., Wold, L. C., & Stange, E. (2014). Monitoring visitors to natural areas in wintertime: Issues in counter accuracy. *Journal of Sustainable Tourism, 22*(4), 550–560.

Blanc, B., Figliozzi, M., & Clifton, K. (2016). How representative of bicycling populations are smartphone application surveys of travel behavior? *Transportation Research Record, 2587*(1), 78–89. https://doi.org/10.3141/2587-10

Bubalo, M., van Zanten, B. T., & Verburg, P. H. (2019). Crowdsourcing geo-information on landscape perceptions and preferences: A review. *Landscape and Urban Planning, 184*, 101–111. https://doi.org/10.1016/j.landurbplan.2019.01.001

Byczek, C., Longaretti, P.-Y., Renaud, J., & Lavorel, S. (2018). Benefits of crowd-sourced GPS information for modelling the recreation ecosystem service. *PLOS ONE, 13*(10), e0202645.

Davies, H., & Dutton, A. (2021). *UK natural capital accounts: 2021*. Office for National Statistics. https://www.ons.gov.uk/economy/environmentalaccounts/bulletins/uknaturalcapitalaccounts/2021.

Day, B. H. (2020). The Value of Greenspace Under Pandemic Lockdown. *Environmental and Resource Economics, 76*(4), 1161–1185. https://doi.org/10.1007/s10640-020-00489-y

Dolan, R., Bullock, J. M., Jones, J. P. G., Athanasiadis, I. N., Martinez-Lopez, J., & Willcock, S. (2021). The flows of nature to people, and of people to nature: applying movement concepts to ecosystem services. *Land, 10*(6), 576. https://doi.org/10.3390/land10060576

Evensen, K. H., Hemsett, G., & Nordh, H. (2021). Developing a place-sensitive tool for park-safety management experiences from green-space managers and female park users in Oslo. *Urban Forestry & Urban Greening, 60*, Article 127057. https://doi.org/10.1016/j.ufug.2021.127057

Fredman, P., Romild, U., Emmelin, L., & Yuan, M. (2009). Non-compliance with on-site data collection in outdoor recreation monitoring. *Visitor Studies, 12*(2), 164–181.

Hansen, A. S. (2021). Understanding recreational landscapes – a review and discussion. *Landscape Research, 46*(1), 128–141. https://doi.org/10.1080/01426397.2020.1833320

Havinga, I., Bogaart, P. W., Hein, L., & Tuia, D. (2020). Defining and spatially modelling cultural ecosystem services using crowdsourced data. *Ecosystem Services, 43*, Article 101091. https://doi.org/10.1016/j.ecoser.2020.101091

Heesch, K. C., Langdon, M., Heesch, K. C., & Langdon, M. (2016). 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. https://doi.org/10.1071/HE16032

Hermes, J., Van Berkel, D., Burkhard, B., Plieninger, T., Fagerholm, N., von Haaren, C., & Albert, C. (2018). Assessment and valuation of recreational ecosystem services of landscapes. *Ecosystem Services, 31*, 289–295. https://doi.org/10.1016/j.ecoser.2018.04.011

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

Kantar. (2021). *Osloborgernes bruk av Marka i 2021. Kartlegging av bruk av Marka, barrierer og muligheter for bruk, spørreundersøkelse gjennomført i 2021.* (15.04.2021).

Lee, K., & Sener, I. N. (2019). Understanding potential exposure of bicyclists on roadways to traffic-related air pollution: Findings from El Paso, Texas, using Strava metro data. *International Journal of Environmental Research and Public Health, 16*(3), 371. https://doi.org/10.3390/ijerph16030371

Lee, K., & Sener, I. N. (2020). Strava Metro data for bicycle monitoring: A literature review. *Transport Reviews*, 1–21. https://doi.org/10.1080/01441647.2020.1798558

Milne, D., & Watling, D. (2019). Big data and understanding change in the context of planning transport systems. *Journal of Transport Geography, 76*, 235–244.

Nelson, T., Ferster, C., Laberee, K., Fuller, D., & Winters, M. (2021). Crowdsourced data for bicycling research and practice. *Transport Reviews, 41*(1), 97–114. https://doi.org/10.1080/01441647.2020.1806943

Niu, H., & Silva, E. A. (2020). Crowdsourced data mining for urban activity: Review of data sources, applications, and methods. *Journal of Urban Planning and Development, 146*(2), 04020007. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000566

Nordh, H., & Østby, K. (2013). Pocket parks for people – A study of park design and use. *Urban Forestry & Urban Greening, 12*(1), 12–17. https://doi.org/10.1016/j.ufug.2012.11.003

Norman, P., Pickering, C. M., & Castley, G. (2019). What can volunteered geographic information tell us about the different ways mountain bikers, runners and walkers use urban reserves? *Landscape and Urban Planning, 185*, 180–190. https://doi.org/10.1016/j.landurbplan.2019.02.015

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

Samuelsson, K., Barthel, S., Colding, J., Macassa, G., & Giusti, M. (2020). *Urban nature as a source of resilience during social distancing amidst the coronavirus pandemic*.

Statistikkbanken. (2022). *Population statistics: Annually, estimated figures*. Statistikkbanken. https://www.ssb.no/en/befolkning/statistikker/folkemengde/aar-berekna.

Strava. (2022). Strava's Year In Sport 2021 charts trajectory of ongoing sports boom. *Strava*. https://blog.strava.com/press/yis2021/.

Sun, Y., Du, Y., Wang, Y., & Zhuang, L. (2017). Examining associations of environmental characteristics with recreational cycling behaviour by street-level Strava data. *International Journal of Environmental Research and Public Health, 14*(6), 644. https://doi.org/10.3390/ijerph14060644

Sun, Y., & Mobasheri, A. (2017). Utilizing crowdsourced data for studies of cycling and air pollution exposure: A case study using Strava data. *International Journal of Environmental Research and Public Health, 14*(3), 274. https://doi.org/10.3390/ijerph14030274

Thorsen, N. H., Bischof, R., Mattisson, J., Hofmeester, T. R., Linnell, J. D. C., & Odden, J. (2022). Smartphone app reveals that lynx avoid human recreationists on local scale, but not home range scale. *Scientific Reports, 12*(1), 4787. https://doi.org/10.1038/s41598-022-08468-7

Venter, Z. S., Barton, D. N., Gundersen, V., Figari, H., & Nowell, M. (2020). Urban nature in a time of crisis: Recreational use of green space increases during the COVID-19 outbreak in Oslo, Norway. *Environmental Research Letters, 15*(10), Article 104075. https://doi.org/10.1088/1748-9326/abb396

Venter, Z. S., Barton, D. N., Gundersen, V., Figari, H., & Nowell, M. S. (2021). Back to nature: Norwegians sustain increased recreational use of urban green space months after the COVID-19 outbreak. *Landscape and Urban Planning, 214*, Article 104175. https://doi.org/10.1016/j.landurbplan.2021.104175

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.