Strava Bias Literature Review

Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists

by Avipsa Roy 1,\*ORCID,Trisalyn A. Nelson 1,A. Stewart Fotheringham 1 andMeghan Winters 2

* The article proposes a method to correct bias in crowdsourced bicycle ridership data from Strava, a fitness app, by using official counts and geographical covariates.
* The article uses data from Maricopa County, Arizona, USA, where Strava data are biased towards recreational riders and do not represent all bicyclists.
* The article applies a three-step approach: (1) quantifying the relationship between Strava and official counts, (2) selecting the most significant geographical variables using LASSO, and (3) fitting a Poisson regression model to predict bias-corrected ridership.
* The article finds that distance to residential areas, distance to green spaces, percentage of white population, median household income, traffic speed, and Strava counts are the key variables for correcting bias.
* The article evaluates the model accuracy using independent data from Tempe, a city within Maricopa County, and shows that it can predict ridership within a reasonable margin of error for most street segments.

THE POTENTIAL OF STRAVA DATA TO CONTRIBUTE IN NON-MOTORISED TRANSPORT (NMT) PLANNING IN JOHANNESBURG M. K. Selalaa, \*, W. Musakwab a Faculty of Engineering and Built Environment. Dept. Quality and Operations Management, University of Johannesburg, Johannesburg-kadibetsos7@gmail.com b Dept. Town and Regional Planning, University of Johannesburg, Johannebsurg- ([wmusakwa@uj.ac.za](mailto:wmusakwa@uj.ac.za))

* The article explores the potential of Strava data, a geolocation based service, to provide information on cycling patterns and trends in Johannesburg, South Africa, where there is limited research on non-motorised transport (NMT).
* The article uses data from Strava Metro, which aggregates GPS tracked activities from Strava users, to analyse the origin and destination, type, time, and frequency of cycling trips in Johannesburg for the year 2014.
* The article reveals that most cycling trips are recreational rather than commuting, and that they vary by season, month, and hour of the day. It also shows that cycling activities are concentrated in the suburbs at the centre and towards the north and east of the city, where there are more high-income households, gated communities, and cycling infrastructure.
* The article acknowledges that Strava data are biased towards certain segments of the population, such as smartphone users, recreational cyclists, and high-income earners, and that they do not capture the needs and preferences of all NMT users, especially the poor and stranded. The article explains that Strava data are not representative of the entire population of cyclists, as they only capture the activities of those who use the Strava app, which requires a smartphone and a certain level of technological literacy. Therefore, Strava data may exclude or underrepresented some groups of people, such as low-income earners, illiterate people, commuters, and those who cycle for necessity rather than choice. The article also notes that Strava data do not provide information on the reasons, motivations, and barriers for cycling, which are important for understanding the needs and preferences of NMT users. The article argues that these biases and gaps in Strava data may affect the accuracy and validity of the analysis and the implications for planning and policy making for NMT in Johannesburg. The article suggests that Strava data should be used with caution and supplemented with other sources of data, such as surveys, interviews, or official counts, to obtain a more comprehensive and reliable picture of the cycling situation in the city.
* The article suggests that Strava data can be a useful tool for planning and policy making for NMT in Johannesburg, as it provides insights into the spatial distribution and behaviour of cyclists. It also recommends promoting the use of NMT and Strava among different groups of people, improving the safety and accessibility of NMT infrastructure, and using other sources of data to complement Strava.

Understanding Potential Exposure of Bicyclists on Roadways to Traffic-Related Air Pollution: Findings from El Paso, Texas, Using Strava Metro Data

by Kyuhyun Lee 1ORCID andIpek N. Sener 2,\*ORCID

* The article examines the potential exposure of bicyclists to traffic-related air pollution in El Paso, Texas, using Strava Metro data that reveal bicycle patterns across the city network.
* The article finds that bicycle volume is significantly associated with higher levels of PM 2.5 emissions and more frequent bus services, implying adverse health effects of exposure to traffic emissions.
* The article also finds that bicycle volume is influenced by various environmental characteristics, such as roadway, bicycle infrastructure, topography, and demographics.
* The article acknowledges the sample bias of Strava users, who tend to be more experienced, recreational, and male riders, and suggests developing data fusion techniques to supplement the bias and remove the uncertainty in representing the total population.
* The article provides insights for policy makers to design safer, healthier, and better adopted bicycle networks and to evaluate the net health impacts of bicycling.

Bias and precision of crowdsourced recreational activity data from Strava

Author links open overlay panelZander S. Venter, Vegard Gundersen, Samantha L. Scott, David N. Barton

* The authors use seven fixed-point counter stations that record the number of people passing by per hour along selected trail segments in Oslo and its surroundings. They extract the daily counts for the time period between 2016 and 2020 which coincided with the availability of the Strava data[1](https://edgeservices.bing.com/edgesvc/chat?udsframed=1&form=SHORUN&clientscopes=chat,noheader,channelstable,&shellsig=7dca9161266c556efe4d4ee6073d7755cc376560&setlang=en-GB&lightschemeovr=1#sjevt%7CDiscover.Chat.SydneyClickPageCitation%7Cadpclick%7C0%7Cc6c03456-abbb-45c6-bdc1-7a5425ffdc90). They also use two counter stations that are collocated with their observation and questionnaire surveys.
* The authors perform in-situ systematic moment observations of recreationists at six locations in Oslo. They count all humans passing by and record their gender, age, and whether they were on a bicycle or not. They also deploy anonymous on-site questionnaire surveys at the same locations to collect information on GPS activity monitoring app usage, income bracket and education level.
* The authors use Strava Metro Service to access Strava data. They download activity counts, stratified by user type and demographic, for the days and trail segments that coincide with their counter stations and surveys. They use only leisure trips and exclude commute trips from their analysis.
* The authors regress observed (i.e. counter station) on Strava activity counts and calculate the linear regression R2 values to quantify the precision of Strava data[3](https://edgeservices.bing.com/edgesvc/chat?udsframed=1&form=SHORUN&clientscopes=chat,noheader,channelstable,&shellsig=7dca9161266c556efe4d4ee6073d7755cc376560&setlang=en-GB&lightschemeovr=1#sjevt%7CDiscover.Chat.SydneyClickPageCitation%7Cadpclick%7C2%7Cc6c03456-abbb-45c6-bdc1-7a5425ffdc90). They also calculate the percentage of observed recreational activities constituted of Strava activities to quantify the bias of Strava data. They do this for temporally and spatially aggregated activity counts, during summer and winter, and for different levels of temporal aggregation. They also compare the percentage composition of Strava and observed recreational users across activity types, gender, age, income and education groups.
* The authors find that the precision with which Strava data captured the spatial variation (R2 = 0.9) was relatively high for monthly time series during summer, although precision degraded at weekly and daily resolutions and during winter. The temporal variation in activity during winter months was less correlated to Strava than during summer months (R2 = 0.51).
* The authors find that 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. However, averaged over the study period, Strava activities represented 3.9 % (±1.1 %) of the recreation activities observed by counter stations.
* The authors find that 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). The largest bias is amongst teenagers and children (<19 years) with Strava data including 23.3 % less activities than observed in this age bracket. The authors also find that Strava pedestrians that were able to complete a questionnaire survey (>19 years) were biased to higher income brackets and education levels.

Leveraging the spatial-temporal resolution of crowdsourced cycling data to improve the estimation of hourly bicycle volume

Author links open overlay panelValerian Kwigizile a, Keneth Morgan Kwayu b, Jun-Seok Oh

* The study selected 19 sites in two cities based on land use, roadway type and bicycle facility type, and collected one-week data of bicycle counts using video cameras and/or pneumatic tubes.
* The study used Strava counts as one of the predictor variables in different probabilistic and Machine Learning models, such as Negative Binomial, Random Forest, K-Nearest Neighbors, Support Vector Machines and Artificial Neural Networks.
* The study also used other predictor variables such as census data, weather data, bicycle facility type, land use type and time of the day to account for the potential biases associated with Strava data.
* The study compared the model performances based on root mean square error (RMSE) and R-squared values, and found that Random Forest and Artificial Neural Network had the best prediction capability.
* The study conducted a simulation study to assess the change in model performance based on different simulated Strava penetration rates, and found that a penetration rate of about 40 percent would improve prediction capability to 96 percent.
* The study developed a web-based tool that can be used to estimate bicycle counts by incorporating crowdsourced data using the Random Forest model. The tool allows the user to input data using a template and produces the results automatically.

strava data have been reported to have inherent sample bias towards cyclists who are recreational riders more than commuters or utilitarian riders. Also, Strava data has been found to be skewed towards male cyclists. Demographic information collected using RiderLog- a crowdsourcing cycling app, was biased towards urban populations. Crowdsourced data can be highly biased towards communities with certain socio-demographic characteristics. For example, in investigating bike sharing program, it was found that the proposed bike sharing stations via crowdsourcing were biased towards white populations. [In addition, cycling activity data collected from smartphone applications underrepresents females, older adults and low-income population1](https://findingspress.org/article/16710-comparing-spatial-associations-of-commuting-versus-recreational-ridership-captured-by-the-strava-fitness-app). The study used other predictor variables such as census data, weather data, bicycle facility type, land use type and time of the day to account for the potential biases associated with Strava data. The study conducted a simulation study to assess the change in model performance based on different simulated Strava penetration rates. [The study found that a penetration rate of about 40 percent would improve prediction capability to 96 percent1](https://findingspress.org/article/16710-comparing-spatial-associations-of-commuting-versus-recreational-ridership-captured-by-the-strava-fitness-app)

Leveraging the spatial-temporal resolution of crowdsourced cycling data to improve the estimation of hourly bicycle volume

Author links open overlay panelValerian Kwigizile a, Keneth Morgan Kwayu b, Jun-Seok Oh c

* The web page context is a research article that explores the potential of using crowdsourced data from fitness apps to improve the estimation of bicycle volume.
* The article uses different probabilistic and machine learning models to test the prediction capability of crowdsourced data from Strava, along with other predictors such as time, weather, and infrastructure variables.
* The article finds that both Random Forest and Artificial Neural Network models have better prediction capability than other models, and that adding crowdsourced data improves the model significantly.
* The article also conducts a simulation study to assess the change in model performance based on different simulated penetration rates of Strava users, and finds that a penetration rate of about 40 percent would improve prediction capability to 96 percent.
* The article develops a web-based tool that can be used to estimate bicycle counts using crowdsourced data and other predictors, and demonstrates its practical application for planners and engineers.
* Crowdsourced data collected using fitness apps have the potential to supplement other data collected through traditional methods to provide spatial and temporal details needed for estimating bicycle exposure.
* However, further research on how to integrate crowdsourced data with traditional data is lacking, and understanding opportunities and limitations associated with crowdsourced data is necessary to guide integration of the data.
* One major limitation of crowdsourced data is the potential sample bias since those being counted have to opt-in to the program and have to own a smartphone and remember to use the app on each trip.
* The data obtained from crowdsourced data can be selective, ultimately introducing a bias when used by planners who are striving to make an equitable distribution of resources and services.
* Strava data have been reported to have inherent sample bias towards cyclists who are recreational riders more than commuters or utilitarian riders, and have been found to be skewed towards male cyclists.
* Demographic information collected using RiderLog, another crowdsourcing cycling app, was biased towards urban populations, which aligns with a study that found that crowdsourced data can be highly biased towards communities with certain socio-demographic characteristics.
* Cycling activity data collected from smartphone applications underrepresents females, older adults and low-income population.
* An intercept survey of 320 cyclists was conducted in the city of Ann Arbor and Grand Rapids located in Michigan, United States as part of the current study, which found that the Strava app is likely to be utilized more by male cyclists and recreational riders than commuters or utilitarian riders.

Using Twitter data for demographic research

Dilek Yildiz1 Jo Munson2 Agnese Vitali2 Ramine Tinati2 Jennifer A. Holland3

* The authors evaluate a series of log-linear models with offsets, measuring the degree to which the models can calibrate the Twitter users’ data set.
* These models are in two broad groups. The first group includes four models that reweight the Twitter sample to match the MYEs total population size or the marginal totals by age group, sex, or location.
* The second group includes three models that reweight the sample to match two sets of marginal totals by age group, sex, and location separately.
* The authors compare the accuracy of the two approaches for estimating age and sex of Twitter users (crowdsourcing and image recognition) using these models.
* The authors aim to identify an improved version of the Total Model, which takes into account the fact that the Twitter sample differs from the ground truth data in terms of association structures between age, sex, and location.

Demographic research with non-representative internet data

Emilio Zagheni

Department of Sociology, University of Washington,

Seattle, Washington, USA, and

Ingmar Weber

Department of Social Computing, Qatar Computing Research Institute,

Doha, Qatar

* The paper reviews the literature and methods for demographic research with non-representative internet data, such as web searches, social media posts, and geo-located records.
* The paper proposes two main approaches to reduce bias and make statistical inference from internet data: calibration and difference in differences.
* Calibration is used when ground truth data are available to estimate the parameters of a model that evaluates the extent of bias in internet data. Difference in differences is used when no ground truth data are available to estimate relative trends over time.
* Calibration is a method that relies on existing traditional data sources and internet data to estimate the bias for specific locations and groups. Once the estimates for the parameters have been generated, the model can be used to make extrapolations for countries or groups for which internet data are available, but no ground truth data about the quantity of interest. The calibration model can also be used to make predictions for future periods, when internet data becomes available before official statistics. This is called “nowcasting”.
* Difference in differences is a method that compares relative changes for a single location or group with relative changes for a reference location or group. This allows the researcher to filter out the bias that affects both locations or groups in a similar way. For example, if the composition of internet users changes over time in both locations or groups, but the underlying quantity of interest changes differently, then the difference in differences method can capture that difference. This method requires assumptions about the functional form of the relation between quantities of interest and their online proxies, as well as internet penetration and socio-demographic variables.
* The paper discusses the potential and limitations of these approaches, and provides examples from migration studies using e-mail and Twitter data.
* The paper aims to stimulate methodological discussions among social scientists who use web data, and to suggest new avenues for research in this area.