Tracks vs. Counters: Towards a Systematic Analysis of Spatiotemporal Factors Influencing Correlation

A. Graser[1,2], P. Stutz[2], M. Loidl[2]

[1] AIT Austrian Institute of Technology, Vienna, Austria
Email: anita.graser@ait.ac.at

[2] University of Salzburg, Salzburg, Austria
Email: {petra.stutz;martin.loidl}@sbg.ac.at

Abstract

Mobility data from tracking apps and stationary counters are often limited by biased sampling, uneven spatial distribution and low spatial coverage, respectively. Data fusion approaches attempt to combine the advantages of both sources. However, the observed correlation of tracks and counts is often mediocre. The reasons for these differences are often ascribed to sampling bias. However, we argue that this is an oversimplification of the real relationship. We present our work in progress concept for relating tracking data and stationary counting data to critically reflect on the factors influencing their correlation and how this can inform data fusion approaches.

1. Introduction

Tracking data opens new opportunities for researchers and practitioners in the mobility domain and particularly in pedestrian and cycling research (Lee and Sener 2020). While stationary counters (such as video, induction loop, radar, infrared) only provide local information, tracking apps can collect quasi-continuous movement data covering the transport network or movement space in general. Stationary counters provide full counts at the respective locations, while the penetration rates of tracking apps are usually far from representative. Therefore, the fusion of tracking and counting data promises to provide additional insights into movement dynamics and spatial variabilities (Romanillos et al. 2016).

The most common way to relate track and counter data is to count the number of tracks at the location of the counter within the same time period. Table 1 presents an overview of recent studies on cycling that bring together track and counter data. The correlation is typically expressed as correlation coefficient R or coefficient of determination R^2. Correlation values vary widely between studies and within studies, between different counting sites and temporal aggregations. However, to the best of our knowledge, the conceptual relation of track data and stationary counting data has so far received no attention within GIScience.

Many studies discuss the non-representative character of crowdsourced data, which are biased towards male users, younger age groups and sportive or leisure cycling (Garber et al. 2019) influenced by users' individual practices, app design, data acquisition setting (Tironi and Valderrama 2017). Consequently, some authors use spatial variables, such as land use and socio-economic environment, to correct skewed distributions of flows (Munira and Sener 2000, Sun et al. 2017).

However, three sources of error are hardly ever considered: First, the spatial deviation of measurements with GNSS-enabled devices and the consequence for spatially relating tracks to stationary counting locations are not taken into account. Second, the quality and suitability of the counting data, which are used as reference, is only scrutinized by Boss et al. (2018), who determine the accuracy of their counting data between +0 and -5%. Third, it remains widely unclear to which degree the spatial distribution of counting stations within the area of

interest, the observation period and the temporal aggregation affects the correlation. We hypothesize that these sources of errors influence the correlation analysis substantially and that this limits the explanatory power of derived conclusions. We therefore aim to provide a concept for critically reflecting studies, which are built upon track and counter data.

Table 1. Studies correlating cycling track and counter data.

Reference	Datasets	Correlation
Hochmair et al. (2019)	Strava & video-based counters	R = 0.55
Jestico et al. (2016)	Strava & manual counts during peak times (7-9am and 3-6pm)	R^2 = 0.40 (hourly) R^2 = 0.56 (peaks) R^2 = 0.58 (total)
Roy et al. (2019)	Strava & municipal counting data	R^2 = 0.76
Conrow et al. (2018)	Strava & manual counts (7-9am)	R = 0.79
Boss et al. (2018)	Strava & permanent counting stations (weekdays)	R = 0.76-0.96* (hourly)
Oksanen et al. (2015)	Sports Tracker & manual counts (workday 7am-7pm)	R^2 = 0.49-0.50* R^2 = 0.72-0.96*[1]
Rupi et al. (2019)	European Cycling Challenge & manual counts (weekdays 8:30-10:30am)	R^2 = 0.73 (hourly)

* depending on counting site; [1] after removal of outliers

2. Influencing Factors

To systematically explore the factors influencing the observed correlation, we identify spatial, temporal, technical and population specific factors, as illustrated by Figure 1. The following sections present preliminary results for a subset of these factors (the ones highlighted in Figure 1) based on track data for Vienna, Austria from the cycle to work campaign "Österreich radelt zur Arbeit" and public counter data from twelve permanent counting stations (Figure 2) for one year (Sept 2015 to Sept 2016), as well as track data from Bike Citizens (a navigation and bicycle community app) and public counter data for Salzburg, Austria. In the future, we will extend our analysis to more cities in order to gain insights into the relation between the two data sources and to critically reflect the presented approach.

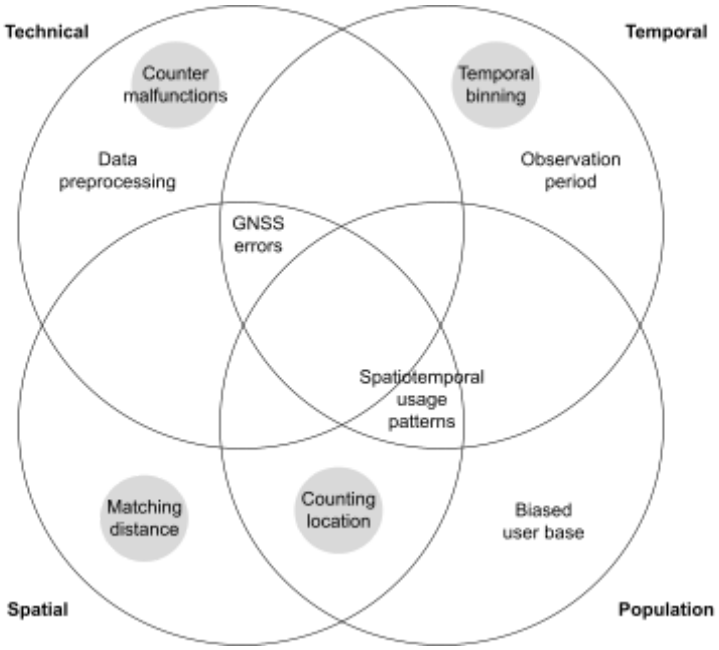

Figure 1. Factors influencing the correlation between tracks and counters.

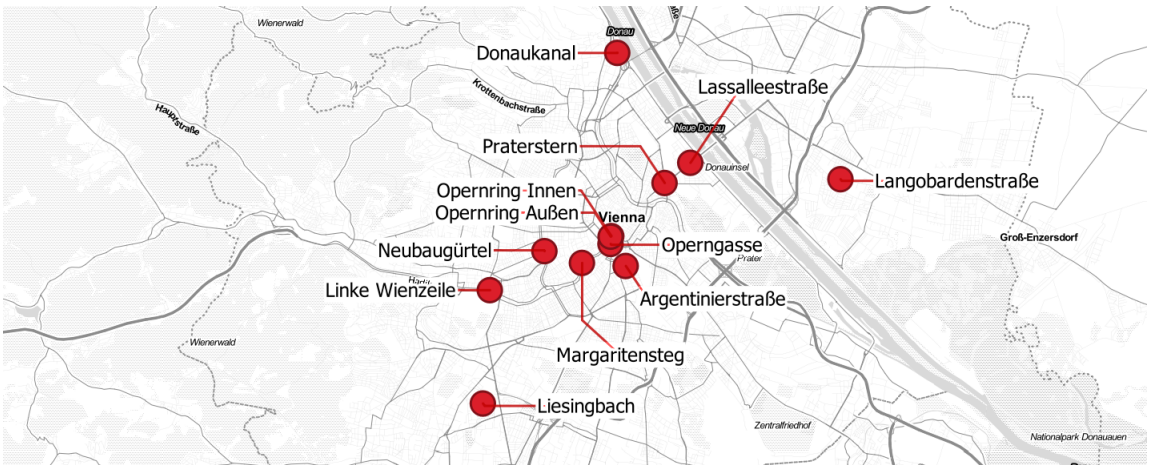

Figure 2. Locations of counters in Vienna. (Background map © Stamen, OpenStreetMap)

2.1 Matching Distance

Depending on the spatial accuracy of the GNSS measurements, tracks are scattered around the true path. Most studies apply a pragmatic approach to match tracks to counting sites based on Euclidean distance. Therefore, the maximum matching distance is key to avoid excluding valid tracks while preventing the inclusion of tracks which did not pass the counter location. We therefore propose an iterative approach using matching distances between 1-70m and a fixed weekly temporal resolution to find the matching distance leading to the highest correlation coefficient R. Our results (Figure 3) show how R varies with the matching distance as well as by station. Two stations (Liesingbach and Langobardenstraße) differ considerably from the other stations, even exhibiting negative correlations. Considering only the ten remaining stations, a matching distance of 30m results in the highest overall correlation and is therefore used for the following analyses.

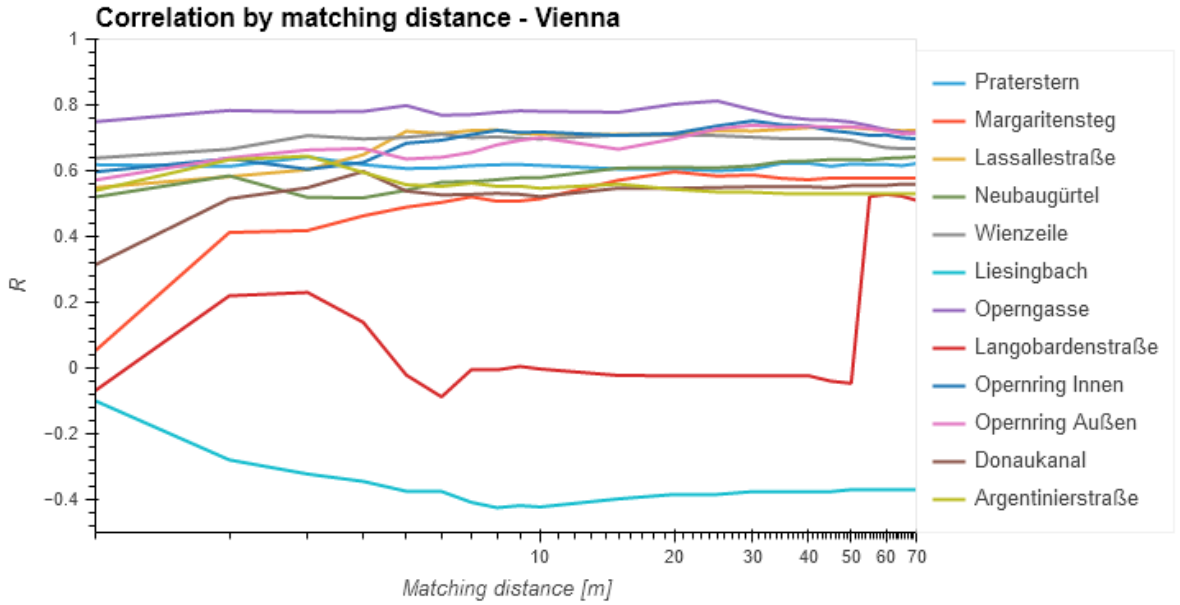

Figure 3. Matching distance effect at the twelve Viennese counting stations.

2.2 Temporal binning

Temporal bin choices affect the correlation since small bins (e.g. hours) lead to random correlations due to uncontrollable external variables in track data. In contrast, large bins result in fewer values, which do not allow for robust correlation analysis. We therefore propose testing a variety of different temporal aggregations and see how those affect the correlation coefficient. We use seven temporal binning modes: monthly (M), weekly (W), and daily (D) over all days or weekdays from Monday to Friday only (Mo-Fr), as well as day of week (DoW) over the whole year. Our results (Figure 4) show that DoW bins outperform all other modes. This confirms the previous statement that aggregation into fewer bins (DoW: 7, M: 12, W: 56, D: 365) tends to result in higher R values.

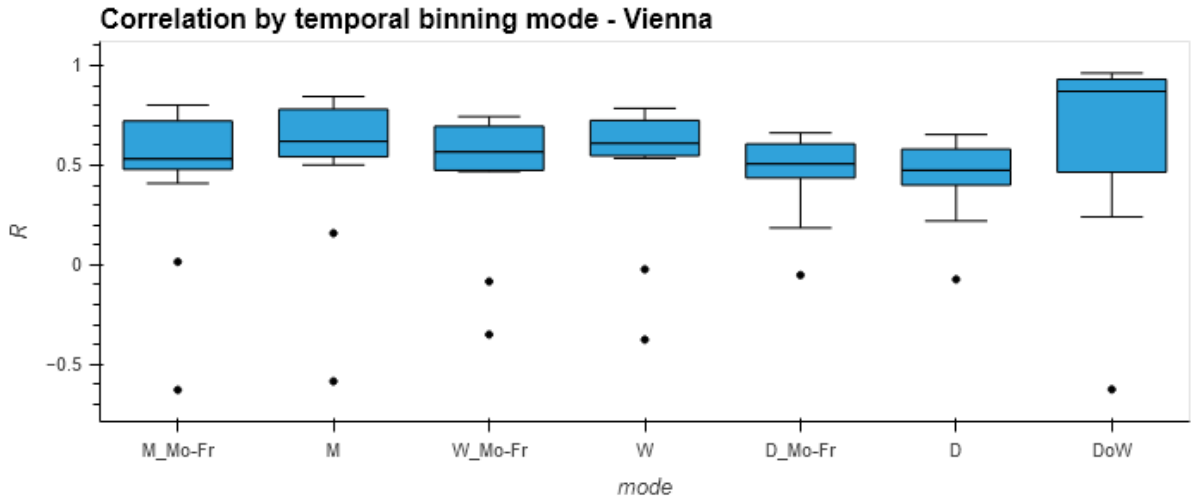

Figure 4. Temporal binning effect at the twelve Viennese counting stations.

2.3 Location

We assume that correlations are higher at central locations with a high volume of cyclists, than in peripheral areas with comparatively low volumes where potential bias has a larger effect. For example, single users who repeatedly travel low-frequency paths will cause an overrepresentation of the respective street segments. We therefore propose to compare the correlation coefficient at different locations. Our results (Figure 5) show a decrease of R in low frequented areas compared to high frequented locations, confirming our assumption that the correlation depends on the location.

2.4 Counter Malfunctions or Reference Data Quality

Data from stationary counters is used as ground truth or reference data. However, counting stations are also prone to errors. We therefore recommend checking for counter malfunctions resulting in abnormally low/high or null counts and exclude corresponding records from further analyses. For illustrating this effect, we used data from the city of Salzburg and compared them to trajectory data from the tracking and navigation app, provided by Bike Citizens. The latter data source is sparse and the overall correlation is weak. However, what becomes evident in Figure 6 is the effect of cleaning the reference data. Interestingly, the changes of the correlation can be in either direction. The counting station “Elisabethkai alt” shows a decreasing correlation after cleaning. This is due to the low number of trajectories at this location and the averaging effect of wrong 0 counts in the reference data.

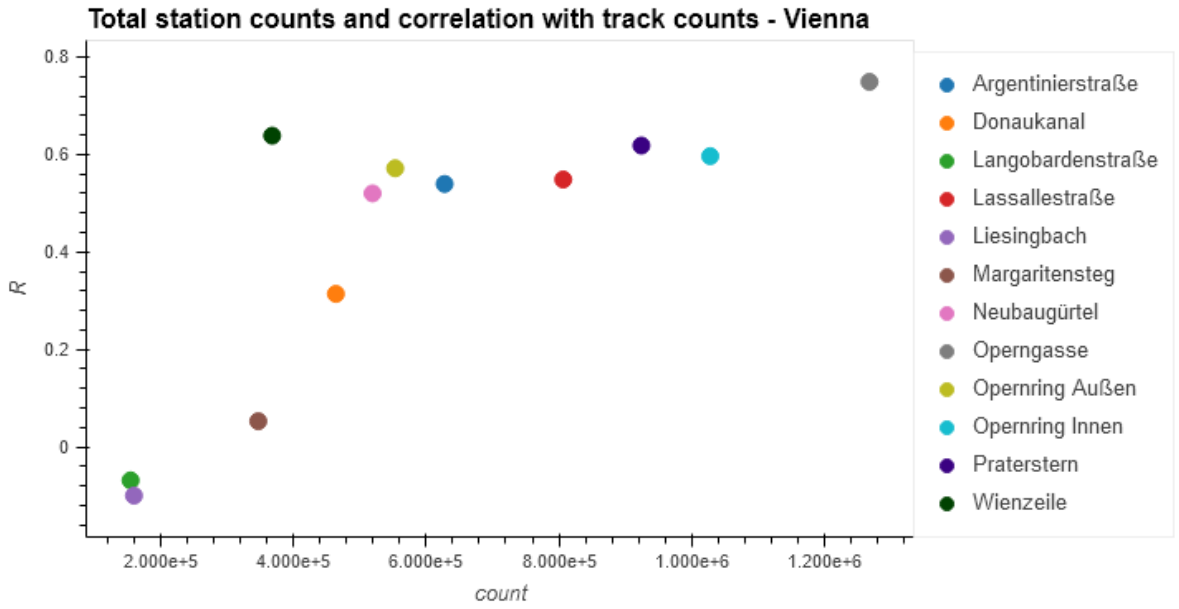

Figure 5. Correlation over total counts at the twelve Viennese counting stations.

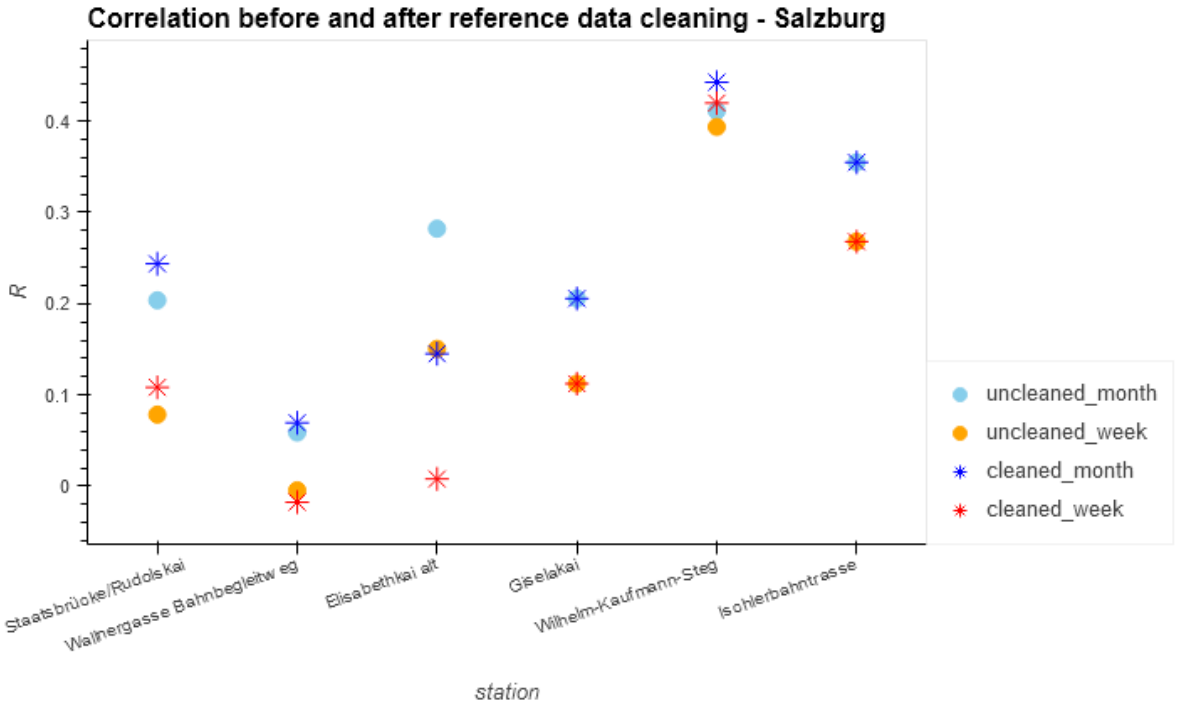

Figure 6. Correlation at six counting stations in Salzburg before and after reference data cleaning, aggregated weekly and monthly.

3. Outlook

Other potential error sources not included in this work in progress paper are the accuracy of data preprocessing steps, including privacy protection measures, cleaning (Loidl et al. 2020) and map matching (Rupi et al. 2019). The observation period may also affect correlation analysis results. Longer time periods may lead to a more robust correlation coefficient. Hence, we plan to compare the correlation coefficient of shorter observation periods with the entire period. Temporal patterns of track data may differ from those of counting stations due to specific behavior patterns of tracking app users. To account for the periodicity at different

temporal scale levels, the patterns over the course of a day (hours per day), over the course of a week (days per week) and over the course of a year (seasons or weeks of a year) should be compared.

Acknowledgements

The Austrian bicycle advocacy group "Radlobby Österreich" provided Bicycle trajectories from the cycle to work campaign "Österreich radelt zur Arbeit" for our analysis in Vienna (Austria). Bike Citizens provided tracking data for the region around Salzburg (Austria). We greatly acknowledge this support.

References

Boss, D., Nelson, T., Winters, M. and Ferster, C. J. 2018. Using Crowdsourced Data to Monitor Change in Spatial Patterns of Bicycle Ridership. Journal of Transport and Health, 9, 226-233.

Conrow, L., Wentz, E., Nelson, T. and Pettit, C. 2018. Comparing Spatial Patterns of Crowdsourced and Conventional Bicycling Datasets. Applied Geography, 92, 21-30.

Garber, M. D., Watkins, K. E. and Kramer, M. R. 2019. Comparing Bicyclists Who Use Smartphone Apps to Record Rides with Those Who Do Not: Implications for Representativeness and Selection Bias. Journal of Transport and Health, 15, 100661.

Hochmair, H. H., Bardin, E. and Ahmouda, A. 2019. Estimating Bicycle Trip Volume for Miami-dade County from Strava Tracking Data. Journal of Transport Geography, 75, 58-69.

Jestico, B., Nelson, T. and Winters, M. 2016. Mapping Ridership Using Crowdsourced Cycling Data. Journal of Transport Geography, 52, 90-97.

Lee, K. and Sener, I. N. 2020. Emerging Data for Pedestrian and Bicycle Monitoring: Sources and Applications. Transportation Research Interdisciplinary Perspectives, 4, 100095.

Loidl, M., Stutz, P., Fernandez La Puente De Battre, M. D., Schmied, C., Reich, B., Bohm, P., Sedlacek, N., Niebauer, J. and Niederseer, D. 2020. Merging Self-reported with Technically Sensed Data for Tracking Mobility Behaviour in a Naturalistic Intervention Study. Insights from the Gismo Study. Scandinavian Journal of Medicine & Science in Sports, 30, 41-49.

Munira, S. and Sener, I.N., 2020. A geographically weighted regression model to examine the spatial variation of the socioeconomic and land-use factors associated with Strava bike activity in Austin, Texas. Journal of Transport Geography, 88, p.102865.

Oksanen, J., Bergman, C., Sainio, J. and 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.

Romanillos, G., Zaltz Austwick, M., Ettema, D. and De Kruijf, J. 2016. Big Data and Cycling. Transport Reviews, 36, 114-133.

Roy, A., Nelson, T. A., Fotheringham, A. S. and Winters, M. 2019. Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists. Urban Science, 3, 62.

Rupi, F., Poliziani, C. and Schweizer, J. 2019. Data-driven Bicycle Network Analysis Based on Traditional Counting Methods and GPS Traces from Smartphone. ISPRS International Journal of Geo-information, 8.

Sun, Y., Du, Y., Wang, Y. and 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), p.644.

Tironi, M. and Valderrama, M. 2017. Unpacking a Citizen Self-tracking Device: Smartness and Idiocy in the Accumulation of Cycling Mobility Data. Environment and Planning D: Society and Space, 36, 294-312.