AGILE: GIScience Series, 7, 40, 2026. https://doi.org/10.5194/agile-giss-7-40-2026
Proceedings of the 29th AGILE Conference on Geographic Information Science, 16–19 June 2026.
Eds.: Evelyn Uuemaa, Alexander Kmoch, Holger Virro, and Simon Scheider
This contribution underwent peer review based on a full paper submission.

Locally Calibrated OpenStreetMap GPS Traces for Traffic Volume Estimation: A Three-Region Benchmark

Michael Schötz [1,2], Martina Schöll [1], and Thomas Limbrunner [1]

[1]Deggendorf Institute of Technology, Deggendorf, Germany
[2]University of Salzburg, Salzburg, Austria

Correspondence: Michael Schötz, michael.schoetz@th-deg.de

Abstract. Traffic-volume estimates are needed far beyond the limited set of roads covered by permanent or periodic count stations. This paper tests whether the publicly available OpenStreetMap (OSM) Planet GPX archive can help fill that gap. We map-match 676 million OSM GPS trackpoints from 2013 to road segments in the United States, Germany, and a UK trace extract, calibrate a length-normalised trace-density model against annual average daily traffic (AADT) counts, and compare its performance with both an attribute-only OSM baseline and an independent global AADT estimate.

The trace signal is present but limited. At the 25 km calibration scale, agreement with the independent AADT estimate is moderate to strong in Germany and Great Britain ($r = 0.78$ and 0.89) but weaker in the US ($r = 0.56$); per-segment accuracy is much weaker ($r = 0.12$–0.43; typical errors of roughly a factor of four or more). OSM road attributes alone already explain most count-station variation ($R^2 = 0.72$–0.79), and trace density adds only 0.6–6.3 percentage points. Timestamped traces also do not recover reliable hour-of-day traffic profiles: their daily shape is dominated by contributor upload behaviour rather than vehicle movement. OSM GPX traces can therefore support coarse consistency checks or serve as a small auxiliary feature where trace coverage and count networks are dense, but they are not a stand-alone substitute for calibrated traffic counts or attribute-based AADT models.

Submission Type. analysis

BoK Concepts. [AM10] Data Mining; [AM11] Network Analysis; [GD4] Data Quality, Metadata and Data Infrastructure

Keywords. crowdsourced data; OpenStreetMap; GPS traces; traffic volume estimation; AADT; map-matching; local calibration

1 Introduction

Traffic volume is a core input to transportation planning, emissions inventories, and road-safety analysis. Yet national monitoring programmes observe only part of the road network. The US HPMS reports AADT for the federal-aid system, about 24 % of US road mileage, using a mixture of counted and modelled state submissions (Federal Highway Administration, 2026, 2019). The German BASt *Dauerzählstellen* mainly cover motorways and federal trunk roads (Fitschen and Nordmann, 2014), and the DfT AADF for Great Britain, while denser, is still built from discrete count points (Department for Transport, 2014).

Crowdsourced OSM GPS traces are a possible additional source because they record contributed movement on roads rather than static OSM road attributes such as class, lane count, or speed tags. Like other volunteered geographic information, however, OSM traces are spatially uneven and shaped by contributor behaviour (Zhang and Zhu, 2018; Haklay, 2010; Neis and Zielstra, 2014). A usable traffic-volume signal would therefore require local calibration; analogous scaling has been shown for crowdsourced bicycle data (Jestico et al., 2016; Dadashova et al., 2020; Roy et al., 2019), but vehicle-volume studies usually depend on commercial or access-restricted probe datasets. It remains unclear whether the public OSM Planet GPX archive can support a similar AADT estimation approach.

We examine this question with a three-region benchmark. We calibrate OSM-GPX trace density against count data in the US, Germany, and Great Britain and evaluate per-way estimates; measure how much trace density adds beyond a tag-only OSM attribute baseline, using the movement-data-free global GRIP4/QRF AADT estimate by van Strien and Grêt-Regamey (2024) as an independent comparator; and test whether OSM trace timestamps contain a usable hour-of-day traffic signal.

We use the 2013-04-09 OSM Planet GPX dump together with same-year AADT data from HPMS (US), BASt (DE),

and DfT AADF (Great Britain). We harmonise roads to three tiers (motorway / arterial / local), map-match trackpoints to OSM ways, and calibrate a length-normalised trace-density model on a 25 km grid. We use cell-scale results as calibration diagnostics, validate applied per-way estimates at segment scale, compare trace density with OSM attribute models, and test hourly profiles against independent count networks. Road geometry for matching comes from March 2026 Geofabrik extracts.

2 Data

OSM GPS traces. We use the publicly released OSM Planet GPX dump of 2013-04-09 (`gpx-planet-2013-04-09.tar.xz`, 21 GB compressed). This dump contains all trace files that contributors have flagged as *public* or *identifiable*. We extract trackpoints with a custom Python pipeline (`scripts/extract_planet_gpx.py`) and assign them to countries using Natural Earth 1:110m polygons. The extractor parsed 2.676 billion trackpoints globally; our study area contains 83.8 M in the continental US, 411 M in Germany, and 181 M in the UK trace extract (676 M total; Fig. 1).

Ground truth. We pair the traces with AADT data from 2013. For the US, we use HPMS shapefiles for the 48 contiguous states and the District of Columbia (columns `aadt_vn` / `f_system_v`). For Germany, we use BASt *Dauerzählstellen Jawe* records on *Autobahnen* (A) and *Bundesstraßen* (B); 1,399 records provide valid 2013 annual DTV values. For Great Britain, we use DfT AADF-by-direction records, with 22,706 unique count points yielding 43,082 directional records. The OSM trace extraction uses the UK country polygon, but DfT count-backed calibration and validation exclude Northern Ireland.

Harmonisation. We map road classes to {motorway, arterial, local} per Table 1. By construction the DE *local* tier is empty (BASt Jawe measures motorway and federal-trunk only). DfT Trunk A-road records (TA), although administratively classified as A-roads in the British numbering scheme, sit physically on roads that OpenStreetMap tags `highway=trunk`, which our OSM harmonisation places in the motorway tier; we therefore assign TA to motorway to match the OSM-side trace counts.

Table 1. Road-class harmonisation across the three ground-truth datasets. Empty cells indicate classes that are not measured by that programme.

Harmonised	US (HPMS)	DE (BASt)	GB (DfT)
motorway	f-system 1–2	A	PM, TM, TA
arterial	f-system 3–4	B	PA, MB
local	f-system 5–7	—	MCU

Equal-area grid. Each study area is projected to its respective equal-area CRS[1] and gridded on aligned 25 km cells whose bounds match the ground-truth extent. Trackpoints are binned to cells with a vectorised pyproj + `np.bincount` pipeline. Cells with zero AADT records or zero traces are dropped.

Data availability. Planet dump: https://planet.openstreetmap.org/gps/. HPMS: https://www.fhwa.dot.gov/policyinformation/hpms.cfm. BASt: https://www.bast.de/DE/Themen/Digitales/HF_1/Massnahmen/verkehrszaehlung/zaehl_node.html. DfT: https://roadtraffic.dft.gov.uk/downloads.

3 Methodology

Map-matching. We snap every trackpoint to its nearest OSM drivable way using perpendicular point-to-line-segment distance with a 30 m buffer. The matcher (`gpx-analyzer-core`, an R-tree-indexed Rust crate bound to Python via PyO3) keeps motorway, trunk, primary, secondary, tertiary, unclassified, and residential ways, including link roads for the major classes; service, track, and pedestrian ways are excluded. Each matched trackpoint inherits the harmonised class of its OSM way (Table 1).

The 30 m buffer matches 63–69 % of trackpoints across the three study areas; the median matched distance is 3 m.

Calibration aggregation. Matched trackpoints are first assigned to OSM ways. For calibration, we aggregate matched trace counts and total mapped road length by 25 km cell i and harmonised class j, producing a trace density $\mathrm{traces}_{i,j}/L_{i,j}$. Ground-truth $\mathrm{AADT}_{i,j}$ is the mean count-station AADT for class j in the same cell; for US HPMS line features, we use a length-weighted mean. The cells therefore define local calibration neighbourhoods, not final traffic-volume outputs.

Local-calibration predictor. The calibration decomposes AADT into a per-class scaling factor, a per-cell contributor-density correction, and a length-normalised trace density:

$$\widehat{\mathrm{AADT}}_{i,j} = \alpha_j \cdot \delta_i \cdot \frac{\mathrm{traces}_{i,j}}{L_{i,j}} \tag{1}$$

where α_j is a study-area-wide scaling factor per class, δ_i corrects for uneven contributor coverage in cell i, and $L_{i,j}$ is the total length in kilometres of class-j ways in cell i. Taking logs gives an additive model $\log(\mathrm{AADT}_{i,j}) - \log(\mathrm{traces}_{i,j}/L_{i,j}) = \log\alpha_j + \log\delta_i + \varepsilon_{i,j}$, which we fit by OLS with class and cell indicator variables. We then apply the fitted factors to individual OSM ways as $\widehat{\mathrm{AADT}}(w) = \alpha_{class(w)}\,\delta_{cell(w)}\,\mathrm{traces}(w)/L(w)$. Length normalisation makes the calibration unit-consistent with per-way application.

[1] EPSG:5070 (US), EPSG:3035 (DE), EPSG:27700 (Great Britain).

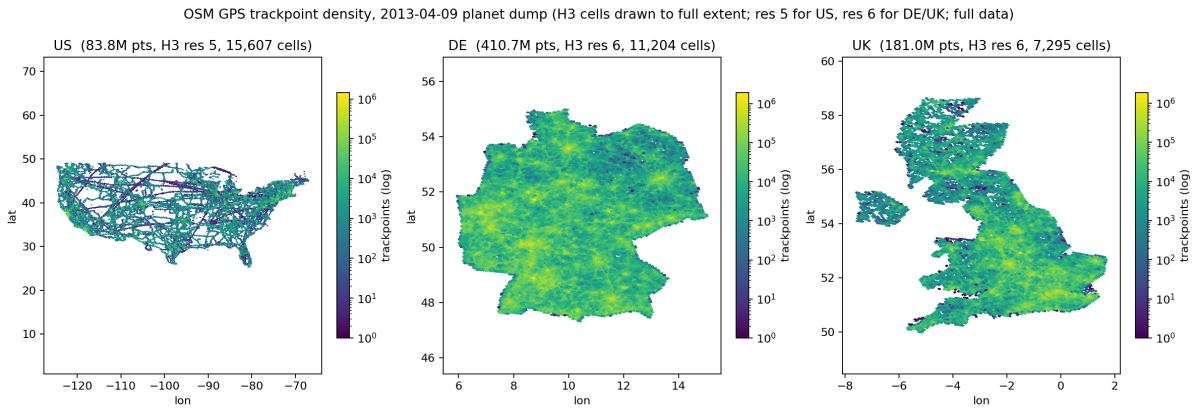

Figure 1. OSM GPS trackpoint density in the 2013-04-09 Planet dump. H3 resolution is 5 for the US (≈253 km^2/cell) and 6 for DE and the UK trace extract (≈36 km^2/cell). Colour encodes trackpoint count per cell (log scale). Contributor concentration follows metropolitan corridors and major roads, with distinct study-area patterns.

Validation. We calibrate at 25 km because finer grids leave too few count stations per cell to estimate δ_i reliably. The fitted model is then applied to individual OSM ways. We report cell-scale R^2 only as a calibration diagnostic and per-segment log-log r against the 2015 layer of van Strien and Grêt-Regamey (2024) as the main validation of per-way estimates. Because this comparator is two years later than the GPX and count data, we treat it as a cross-estimate consistency check rather than same-year ground truth. A continuous spatial smoother for the calibration factor could reduce hard cell-boundary effects, but is not evaluated here.

4 Results

4.1 Within-class correlations after map-matching

The first test is whether map-matched trace density is ordered like traffic volume within each road class. Table 2 shows a positive relationship for motorways in every study area ($r = 0.26$–0.64) and for arterials in Germany and Great Britain ($r = 0.21$–0.34), but the US arterial signal is essentially zero. Local roads in the US and Great Britain are similar to arterials. Germany has no local-road counts because BASt monitors only *Autobahnen* and *Bundesstraßen*.

Table 2. Within-class Pearson correlation between log trace density (traces / road length per cell-class) and $\log(\mathrm{AADT})$ per study area and harmonised class after 30 m perpendicular-buffer map-matching. *n/a*: class not measured by the ground-truth programme.

Class	US	DE	GB
motorway	0.26	0.64	0.28
arterial	0.00	0.34	0.21
local	0.22	n/a	0.33

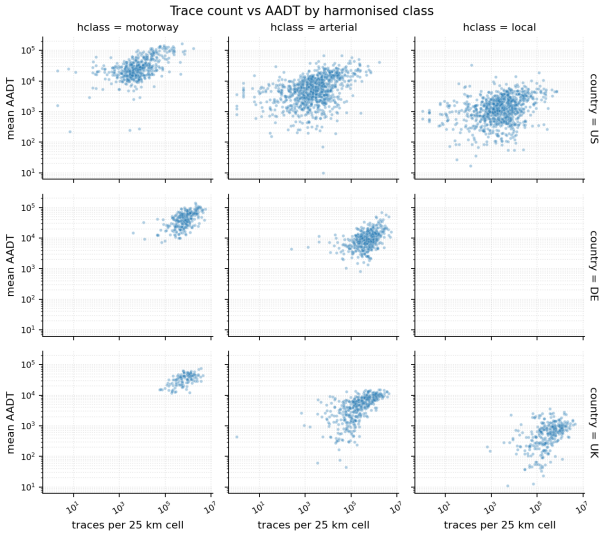

Figure 2. Map-matched trace count vs mean AADT, log–log, per 25 km cell, faceted by study area (rows) and class (columns). Empty panels reflect classes not measured by the ground-truth programme.

4.2 Local-calibration diagnostics

Local calibration improves the cell-scale diagnostic fit, especially where count stations are dense. Figure 3 compares fitted and observed AADT under eq. 1 for the cell-class units used to estimate the calibration factors. These values are not proposed as a deployed traffic-volume map; they test whether the calibration is stable enough to apply to individual ways. In-sample log-space R^2 is 0.86 in Great Britain, 0.86 in Germany, and 0.32 in the US. Under 5-fold cross-validation, held-out R^2 remains positive in Great Britain (0.48) and Germany (0.17) but falls to -1.50 in the US, where trace density is much lower (Sect. 5).

Diagnostic fit follows the density and breadth of the count programme. The DfT AADF for Great Britain covers all

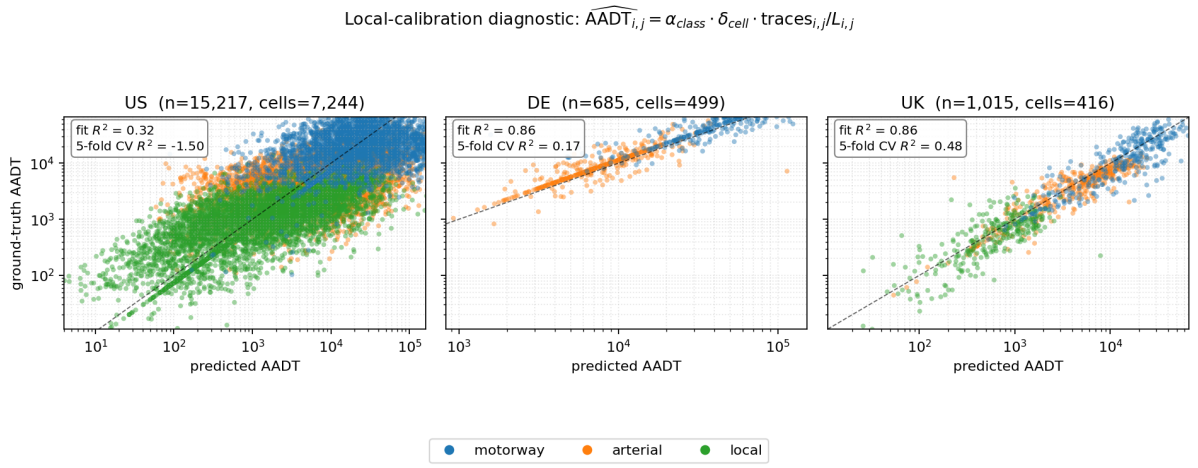

Figure 3. Local-calibration diagnostic, per study area and harmonised class. Points are individual (cell, class) calibration units; dashed line is 1:1. In-sample and 5-fold cross-validated log-space R^2 are annotated in each panel.

three tiers and gives the strongest cross-validated fit. Germany's motorway-plus-federal-trunk coverage supports a strong in-sample fit but weaker held-out prediction because many cells contain only one measured class. The US result is limited by sparse trace coverage, especially outside the highest-volume network.

4.3 Aggregate cross-comparison with the van Strien 2015 AADT estimate

As a second aggregate diagnostic, we compare the trace-based estimates with the 2015 extra-urban AADT predictions of van Strien and Grêt-Regamey (2024) at 25 km cell scale (Table 3). Each GRIP4 line is assigned to a cell by majority vote across five sample points along the line. Raw trace density correlates only moderately with the van Strien layer ($r = 0.37$–0.51). Adding class scaling improves Germany and Great Britain, and the full local-calibration model reaches $r = 0.56$ (US), 0.78 (DE), and 0.89 (GB). The same comparison between ground truth and the van Strien layer gives $r = 0.92$ (US), 0.80 (DE), and 0.93 (GB), which provides a practical ceiling for this cross-estimate comparison. The trace model approaches that ceiling in Germany and Great Britain, but remains weaker in the US.

Table 3. Log-log Pearson correlation with the van Strien and Grêt-Regamey (2024) 2015 AADT layer per 25 km cell. Three trace-based predictors are length-normalised; GT vs VS is the 2013 ground truth compared with the 2015 van Strien layer. Lines are assigned to cells by majority vote across five sample points per line.

Area	n	traces/L	class	local	GT–VS
US	8,312	0.40	0.32	0.56	0.92
DE	684	0.51	0.68	0.78	0.80
GB	701	0.37	0.50	0.89	0.93

4.4 Per-segment validation against van Strien

Aggregate agreement averages over many road segments, so it can overstate the accuracy available to users who need segment-level AADT. We therefore repeat the van Strien comparison at native segment resolution. Each GRIP4 line is sampled at five points and snapped to the nearest OSM way within a 100 m buffer; the line is assigned to the OSM way receiving the most votes. We retain only class-consistent matches (van Strien highway $\rightarrow$ OSM motorway; van Strien primary/secondary $\rightarrow$ OSM arterial) and compare van Strien's 2015 per-segment median AADT with our per-way predictor $\widehat{\text{AADT}}(w) = \alpha_{class(w)} \cdot \delta_{cell(w)} \cdot \text{traces}(w)/L(w)$. This step is the main validation of the trace-based AADT estimate because it evaluates the scale at which road-volume estimates are normally used.

The segment-level comparison is much weaker than the aggregate result. Log-log r is 0.41 in Great Britain, 0.43 in Germany, and 0.12 in the US ($n = 46{,}113$, 73,404, and 75,358, respectively). Log-space RMSE is 1.4–1.9, meaning typical segment predictions differ from the van Strien layer by factors of 4–7. Median ratios are closer: 0.48 (GB), 0.73 (Germany), and 1.30 (US). The very low US r is consistent with the trace-density boundary case discussed in Sect. 5.

4.5 How much signal is distinct from OSM attributes?

The previous results show that trace density carries traffic information, but not whether that information is distinct from ordinary OSM road attributes. We test this by training random forests on count stations with three feature sets: OSM attributes only (harmonised class, highway tag, lanes, maxspeed, oneway, presence of an OSM `ref` tag, length), trace density plus class and length, and the

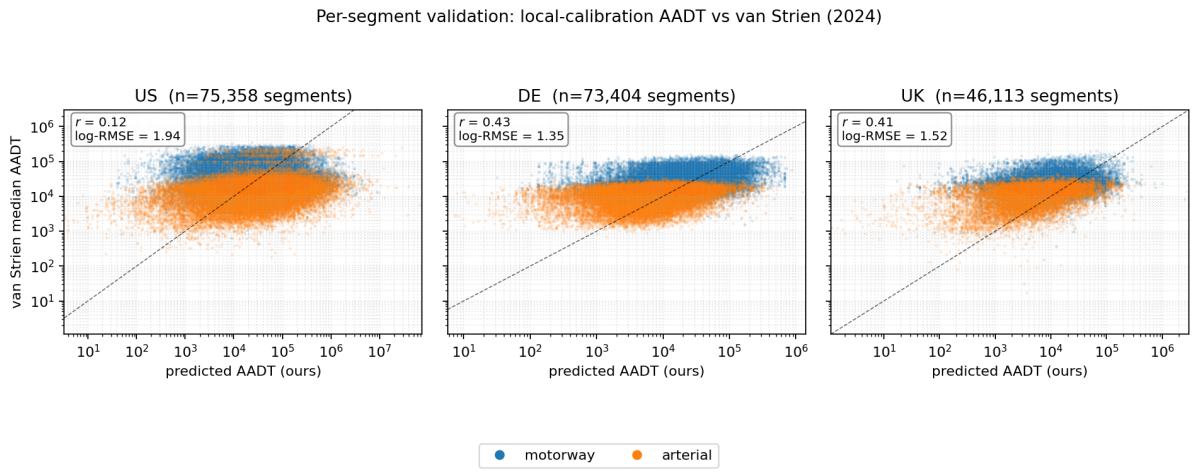

Figure 4. Per-segment validation: length-normalised local-calibration AADT predictions vs van Strien and Grêt-Regamey (2024) 2015 median AADT, log–log, for every OSM way matched to a van Strien GRIP4 line by majority vote across five sample points within a 100 m buffer. Dashed line is 1:1.

union of both. Under 5-fold CV on log-AADT, attributes alone reach $R^2 = 0.727$ in the US ($n = 1{,}000{,}000$, random subsample of national HPMS), 0.789 in Great Britain ($n = 43{,}006$), and 0.717 in Germany ($n = 1{,}397$). Adding trace density raises these scores to 0.732, 0.818, and 0.780, respectively. Thus traces add a small but measurable improvement beyond OSM attributes: +0.6 percentage points in the US, +2.8 in Great Britain, and +6.3 in Germany.

4.6 Temporal resolution: contributor upload bias dominates the trace signal

Finally, we test whether trace timestamps provide usable hourly traffic profiles. We compare trace-derived hour-of-day shapes with two independent hourly count datasets: BASt *Stundenwerte* 2022 for Germany (1,391 stations) and FHWA TMAS 2013 for the US (299 stations with trace coverage out of 6,879 stations with weekday hour-of-day profiles). We compare within-day percentages rather than absolute volumes, since the BASt hourly file we obtained covers 2022 only. For each station, we compare three non-uniform predictors against the station's hourly profile with Pearson r: the class-mean profile from ground truth, the country-mean profile, and the trace-based profile. In Germany, the class-mean profile reaches median $r = 0.98$ and the country mean reaches 0.97, while the trace-based profile reaches only 0.36. In the US, class and country means both reach 0.97, while the trace-based profile reaches 0.25.

Trace timestamps peak at 15–16 local hour at German and US count-station locations (the UK trace extract peaks at 12–13) and miss the morning commute (Fig. 5). The dominant hourly signal in OSM traces is contributor upload behaviour, not vehicle movement. Shin et al. (2025) studied OSM contributor timing differences with survey-linked contribution data; here we quantify a related timestamp bias against national hourly count networks. Recovering an hour-of-day signal from traces would therefore require a contributor-activity correction calibrated against external hourly counts.

5 Discussion and Conclusions

Overall finding. The evidence is mixed. OSM Planet GPX traces contain a traffic-volume signal, and local calibration can recover that signal at coarse spatial scales where both trace coverage and count coverage are dense. However, the public trace archive is not sufficient as a stand-alone AADT estimate. Segment-level accuracy remains modest, the gain over OSM road attributes is small, and timestamp-derived hourly profiles are dominated by contributor behaviour.

The trace signal is mostly redundant with OSM attributes. For AADT magnitude estimation, OSM road tags are strong predictors. Class, lanes, maxspeed, oneway, presence of an OSM `ref` tag, highway tag, and length predict log-AADT with $R^2 = 0.72$–0.79 under random forests on count stations. For context, van Strien and Grêt-Regamey (2024)'s movement-data-free GRIP4/QRF model reports pseudo-$R^2 = 0.74$. Length-normalised trace density adds 0.6 (US), 2.8 (GB), and 6.3 (DE) percentage points. The open trace signal is therefore useful, but mostly as a small supplement to attributes rather than as the primary predictor.

Hourly profiles need separate calibration. The timestamps in the GPX archive should not be interpreted as an hourly traffic sensor without additional correction. In the two-country comparison, trace-based profiles reach $r = 0.36$ in Germany and 0.25 in the US, while simple class-mean baselines from count data reach $r \geq 0.97$ in both. The traces miss the morning commute and peak in mid-afternoon, consistent with upload-time bias rather than vehicle-movement timing. Operational hourly use would

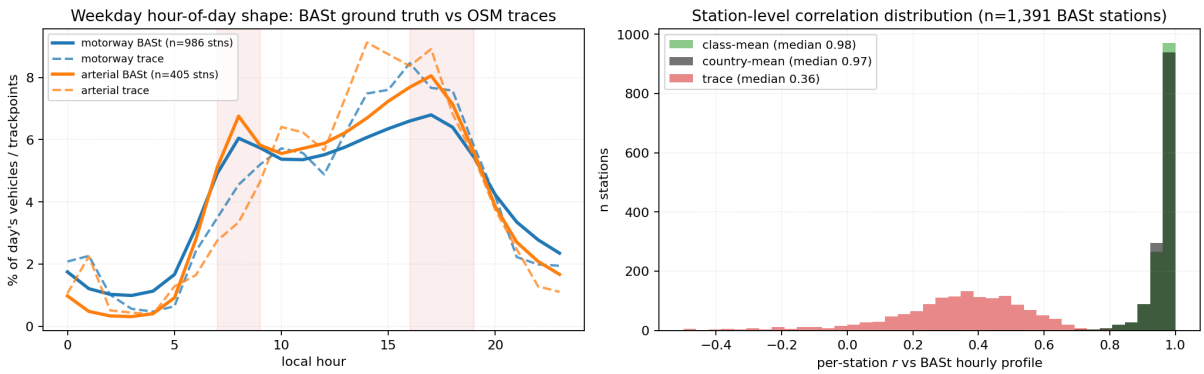

Figure 5. Left: weekday hour-of-day shape from BASt 2022 ground truth (solid) and from 2013 OSM traces matched to each station (dashed), averaged per class. Red bands mark canonical 07–09 and 16–19 commute windows. BASt shows the expected double-peak; traces show a single afternoon bulge, missing the morning commute. Right: per-station r vs BASt, $n = 1{,}391$.

require a contributor-activity correction calibrated against external counts.

Limitations. The 25 km grid is coarser than the scale at which contributor density actually varies, but finer grids require denser count data than most countries provide. A continuous spatial smoother or interpolation of the calibration factor around count stations may be preferable to hard grid cells, especially near cell boundaries; this benchmark does not test that extension. The systematic US weakness is a trace-density boundary case: the 2013 Planet GPX dump has only $\sim$10 pts/km^2 in the continental US, compared with $\sim$742 in the UK trace extract and $\sim$1,151 in Germany. The median valid calibration-unit trace count in the US is 409, vs. 53,206–89,749 in Great Britain and Germany ($\sim$130–219 $\times$ lower); the local δ_i correction becomes noise-dominated, and held-out predictive performance collapses. The calibration therefore generalises only where trace coverage is dense enough; the 2013 US dump does not meet that threshold. Great Britain also has class heterogeneity: some Principal A-road records are harmonised to arterial but snap physically to OSM `highway=trunk` segments, adding residual variance. Finally, calibration constants (α, δ) are year-specific and must be refit for other OSM trace archive snapshots.

Declaration of Generative AI in writing

The authors declare that they have used Generative AI tools in the preparation of this manuscript. Specifically, Claude (Anthropic) was used for literature review and code development for data-processing pipelines. All AI-generated content was reviewed, verified, and edited by the authors, who take full responsibility for the accuracy and integrity of the work.

Acknowledgements

The research leading to these results is funded by the German Federal Ministry for Economic Affairs and Energy within the project “just better DATA - Effiziente und hochgenaue Datenerzeugung für KI-Anwendungen im Bereich autonomes Fahren”. The authors thank the project consortium for its cooperation.

References

Dadashova, B., Griffin, G. P., Das, S., Turner, S., and Sherman, B.: Estimation of Average Annual Daily Bicycle Counts Using Crowdsourced Strava Data, Transportation Research Record, 2674, 390–402, https://doi.org/10.1177/0361198120946016, 2020.

Department for Transport: Road traffic statistics: Annual average daily flows by direction, Great Britain, 2013, https://roadtraffic.dft.gov.uk/downloads, accessed 2026-04-14, 2014.

Federal Highway Administration: Status of the Nation’s Highways, Bridges, and Transit: Conditions and Performance Report to Congress, 23rd Edition, Chapter 1: System Assets, U.S. Department of Transportation, https://www.fhwa.dot.gov/policy/23cpr/chap1.cfm, 2019.

Federal Highway Administration: Highway Performance Monitoring System (HPMS), U.S. Department of Transportation, https://www.fhwa.dot.gov/policyinformation/hpms.cfm, accessed 2026-05-07, 2026.

Fitschen, A. and Nordmann, H.: Verkehrsentwicklung auf Bundesfernstraßen 2013, BASt-Reihe Verkehrstechnik V 244, Bundesanstalt für Straßenwesen (BASt), https://bast.opus.hbz-nrw.de/solrsearch/index/search/searchtype/series/id/7/rows/20/sortfield/year/sortorder/asc/facetNumber_subject/all/facetNumber_year/all/subjectfq/Jahr, annual evaluation of automatic permanent counting stations, 2014.

Haklay, M.: How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets, Environment and Planning B: Planning and Design, 37, 682–703, https://doi.org/10.1068/b35097, 2010.

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

Neis, P. and Zielstra, D.: Recent developments and future trends in volunteered geographic information research: The case of OpenStreetMap, Future Internet, 6, 76–106, https://doi.org/10.3390/fi6010076, 2014.

Roy, A., Nelson, T. A., Fotheringham, A. S., and Winters, M.: Correcting Bias in Crowdsourced Data to Map Bicycle Ridership of All Bicyclists, Urban Science, 3, 62, https://doi.org/10.3390/urbansci3020062, 2019.

Shin, H., Gardner, Z., Solomon, G., and Basiri, A.: Diagnosing Spatial and Temporal Biases of OSM Contributors: Identifying Differences Between Gender and Age from an Online Survey, Annals of the American Association of Geographers, 115, 782–802, https://doi.org/10.1080/24694452.2024.2447507, 2025.

van Strien, M. J. and Grêt-Regamey, A.: A Global Time Series of Traffic Volumes on Extra-Urban Roads, Scientific Data, 11, 470, https://doi.org/10.1038/s41597-024-03287-z, traffic volume data added to GRIP4 road network, 2024.

Zhang, G. and Zhu, A.-X.: The Representativeness and Spatial Bias of Volunteered Geographic Information: A Review, Annals of GIS, 24, 151–162, https://doi.org/10.1080/19475683.2018.1501607, 2018.