Supplementary Material

Global evaluation of Overture Maps reveals source-driven disparities in urban geospatial data

Levente Juhász[1,2,†] and Hartwig H. Hochmair[1,†]

[1] Fort Lauderdale Research and Education Center, School of Forest, Fisheries and Geomatics Sciences, University of Florida/IFAS. Davie, FL, USA
[2] Geospatial Analytics, Technology and Open Research (GATOR) Lab, University of Florida

† These authors contributed equally to this work.

Levente Juhász, levente.juhasz@ufl.edu, ORCID: https://orcid.org/0000-0003-3393-4021
Hartwig H. Hochmair, hhhochmair@ufl.edu, ORCID: https://orcid.org/0000-0002-7064-8238

---

1. Analysis results 2
2. Data download and indexing pipeline 8
2.1. Data download 8
2.2. H3 city assignment index 9
2.3. ISEA4H analysis zone index 9
3. Places taxonomy and specificity score 11
3.1. Taxonomy structure 11
3.2. Specificity score (Sk) 12
4. Building overlap computation (B-03): Batched city-loop architecture 13
4.1. Motivation 13
4.2. City-loop architecture 13
4.3. Memory management, resumability, and batched execution 14
5. Ground-truth sampling design and adaptive zone protocol 14
5.1. Stratification framework 14
5.2. Analysis zone 16
5.3. Ground-truth building completeness and precision 16
6. Binomial GLMM 18
6.1. Model specification 18
6.2. GLMM results 19
6.3. Model validation 22
References 23

1. Analysis results

The full set of analysis tasks, the corresponding quality dimension and coverage are listed in Table S1.1.

Table S1.1. Summary of analysis tasks.

Task	Theme	Description	ISO 19157 quality element / metadata category	Metric(s)	Cities	Year
P-01	Places	Data sources	Lineage (data quality metadata)	Source proportion, Pielou's Evenness (*E*)	11,422	2024, 2026
P-02	Places	Co-location	Positional accuracy	Co-location rate	11,422	2024, 2026
P-03	Places	Primary category quality	Thematic accuracy	Null category rate, specificity score	11,422	2024, 2026
B-01	Buildings	Data sources	Lineage (data quality metadata)	Source proportion, dominant source per city	11,422	2024, 2026
B-02	Buildings	Attribute completeness	Completeness	% buildings with subtype, height, floor count	11,422	2024, 2026
B-03	Buildings	Geometric overlap	Logical consistency, positional accuracy	Building area overlapping another footprint	11,422	2024, 2026
GT-01	Buildings	Manual counting	Completeness	Building completeness and precision per city	338	2026
GT-02	Buildings	Inferential modeling	Completeness	Binomial GLMM: region, city, size, population density, and source composition effects on completeness and precision	320	2026

Table S1.2 extends main Table 1 to the regional level, reporting all five intrinsic POI quality metrics separately for each of the nine macro-regions. Fig. S1.1 maps the city-level POI null category rate for both 2024 and 2026 releases across all 11,422 GHSL urban areas, illustrating the geographic distribution of the improvement that drives the S□ gains shown in Fig. S1.2. Fig. S1.2 maps the city-level mean specificity score (*S*□) for both data releases; regional medians are reported in Table S1.2. Table S1.3 extends main Table 2 to the regional level, reporting building count, dominant source, source evenness, attribute completeness (subtype and height), and geometric overlap separately for each of the macro-regions across both releases. Fig. S1.3 maps the dominant building data source by city, in the two releases, while Fig. S1.4 shows the geometric overlap of buildings by city.

Table S1.2. Overture Places intrinsic quality by world region, 2024 and 2026. Regional summaries are computed across all GHSL-defined urban areas (n = 11,422) grouped into nine world regions following the GC_DEV_USR_2025 classification. All metrics are computed over GHSL urban areas only. Number of POIs shows the total feature count within GHSL urban areas for that region. Numerical cells in P-01 through P-03 report the city-level median with the city-level mean in parentheses; Δ columns report the change in the city-level median between the November 2024 and April 2026 data releases. . P-01: Number of Sources: number of unique data sources contributing POIs per city. P-01: Evenness: Pielou's evenness index E = H / ln(S) ∈ [0, 1]. P-01: Dominant Source: the source contributing the largest share of regional POIs, with that share in brackets. P-02: % Co-located: city-level percentage of POIs sharing an exact coordinate with at least one other feature. P-02: Max Stack: city-level maximum number of POIs sharing an identical coordinate. P-03: % Null Cat.: city-level percentage of POIs with no category assigned. P-03: Specificity S□: city-level mean branch-aware log-normalised specificity score ∈ [0, 1] (see Supplementary Methods §3).

				P-01 (Data Sources)								P-02 (Co-location)						P-03 (Category Quality)					
	Number of POIs			Number of Sources			Evenness			Dominant Source		% Co-located			Max Stack			% Null Category			Specificity (Sk)		
Region	2024	2026	Δ	2024	2026	Δ	2024	2026	Δ	2024	2026	2024	2026	Δ	2024	2026	Δ	2024	2026	Δ	2024	2026	Δ
Australia and New Zealand (ANZ)	635,335	881,110	38.7%	2.0 (2.0)	6.5 (6.2)	4.5	0.63 (0.63)	0.52 (0.53)	-0.11	meta [84.0%]	meta [63.0%]	15.0 (14.7)	10.5 (10.0)	-4.4	46.5 (52.6)	26.5 (40.2)	-20	7.0 (7.1)	4.3 (4.0)	-2.7	0.72 (0.72)	0.81 (0.81)	+0.09 (+0.09)
Central and Southern Asia (CSA)	3,010,448	3,937,652	30.8%	2.0 (1.7)	4.0 (3.7)	2	0.37 (0.33)	0.35 (0.31)	-0.02	meta [88.8%]	meta [85.0%]	39.6 (38.2)	13.6 (13.5)	-26	25.0 (42.2)	10.0 (20.6)	-15	19.1 (20.8)	3.1 (3.8)	-15.9	0.68 (0.68)	0.76 (0.75)	+0.08 (+0.08)
Eastern and South-Eastern Asia (ESEA)	6,309,118	8,822,944	39.8%	1.0 (1.5)	3.0 (2.7)	2	0.0 (0.09)	0.11 (0.37)	0.11	meta [98.1%]	meta [89.5%]	10.4 (17.0)	2.6 (7.0)	-7.8	5.0 (29.2)	2.0 (12.6)	-3	12.5 (14.6)	1.0 (2.9)	-11.5	0.70 (0.73)	0.85 (0.84)	+0.15 (+0.11)
Europe	6,613,519	9,583,896	44.9%	2.0 (1.8)	5.0 (5.0)	3	0.35 (0.38)	0.48 (0.49)	0.13	meta [86.4%]	meta [69.4%]	10.5 (14.4)	8.2 (8.3)	-2.3	40.0 (49.1)	12.0 (17.4)	-28	8.1 (8.8)	3.4 (3.9)	-4.7	0.70 (0.70)	0.79 (0.79)	+0.09 (+0.09)
Latin America and the Caribbean (LAC)	5,908,246	8,662,478	46.6%	2.0 (1.9)	4.0 (3.9)	2	0.14 (0.14)	0.12 (0.12)	-0.03	meta [97.2%]	meta [95.1%]	21.5 (26.2)	5.7 (8.3)	-15.8	52.0 (69.6)	18.0 (31.6)	-34	10.4 (11.5)	3.4 (3.6)	-7	0.68 (0.68)	0.78 (0.76)	+0.10 (+0.08)
Northern Africa and Western Asia (NAWA)	1,743,196	2,485,106	42.6%	2.0 (1.5)	3.0 (2.8)	1	0.03 (0.08)	0.08 (0.14)	0.06	meta [97.7%]	meta [91.6%]	33.6 (32.7)	8.3 (10.1)	-25.4	30.0 (67.3)	9.0 (31.0)	-21	17.8 (18.9)	4.4 (5.7)	-13.5	0.69 (0.68)	0.77 (0.76)	+0.08 (+0.08)
Northern America (NA)	5,815,932	7,078,573	21.7%	2.0 (2.0)	6.0 (6.1)	4	0.74 (0.74)	0.55 (0.56)	-0.19	meta [79.5%]	meta [62.4%]	12.4 (12.8)	9.8 (10.0)	-2.6	52.5 (85.8)	21.0 (39.2)	-31.5	8.1 (8.1)	4.3 (4.4)	-3.8	0.78 (0.77)	0.84 (0.84)	+0.06 (+0.07)
Oceania	11,107	14,460	30.2%	2.0 (1.6)	3.0 (2.9)	1	0.07 (0.09)	0.09 (0.11)	0.03	meta [98.6%]	meta [97.1%]	33.3 (33.3)	17.8 (18.0)	-15.6	38.5 (37.7)	27.5 (30.1)	-11	14.2 (17.3)	3.0 (2.9)	-11.1	0.64 (0.63)	0.74 (0.74)	+0.10 (+0.11)
Sub-Saharan Africa (SSA)	898,218	916,505	2.0%	1.0 (1.4)	2.0 (2.1)	1	0.0 (0.08)	0.09 (0.2)	0.09	meta [96.7%]	meta [89.4%]	60.9 (53.7)	21.1 (22.4)	-39.8	13.0 (36.4)	3.0 (14.6)	-10	22.0 (23.1)	1.8 (4.7)	-20.2	0.65 (0.64)	0.73 (0.72)	+0.08 (+0.08)
Global	30,945,119	42,382,724	37.0%	2.0 (1.6)	4.0 (3.4)	2	0.09 (0.22)	0.24 (0.3)	0.16	meta [90.7%]	meta [80.7%]	27.5 (30.2)	8.7 (11.5)	-18.7	27.0 (46.5)	8.0 (20.3)	-19	14.3 (16.7)	3.1 (3.9)	-11.2	0.69 (0.69)	0.78 (0.78)	+0.09 (+0.09)

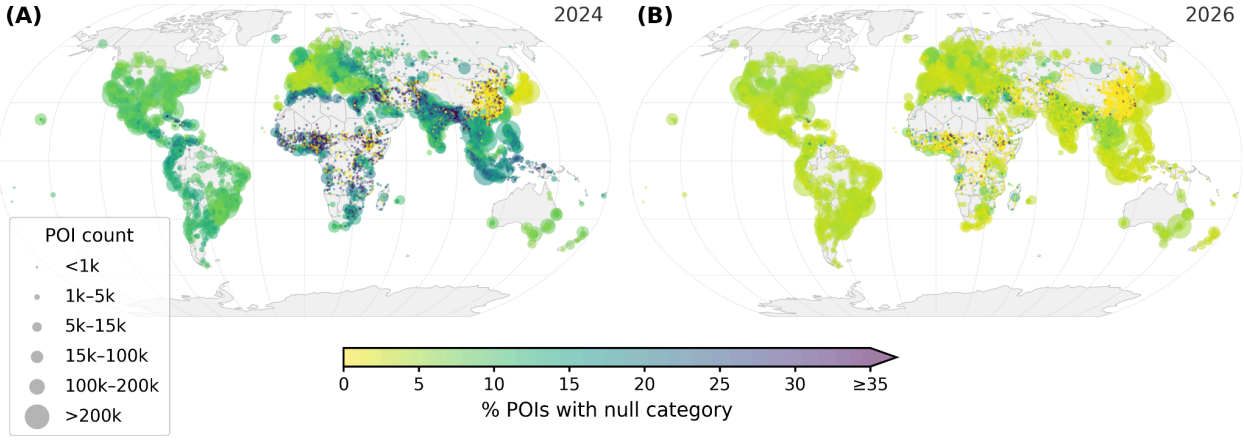


Fig. S1.1. Percentage of uncategorized POIs with null category at the city level, shown for November 2024 (A) and April 2026 (B). Each circle represents one GHSL urban area (*n* = 11,422); dot size is proportional to city-level POI count. Color encodes the percentage of POIs without an assigned category. The 2024 data release (panel A) reflects widespread shares of uncategorized POIs ranging from 10% to over 35%, primarily because Meta-sourced POIs lacked category assignments. The 2026 release (panel B) shows uniform improvement across geographic

regions, with the most pronounced reductions observed in Sub-Saharan Africa (−20.2 percentage points) and Central and Southern Asia (−15.9 pp).

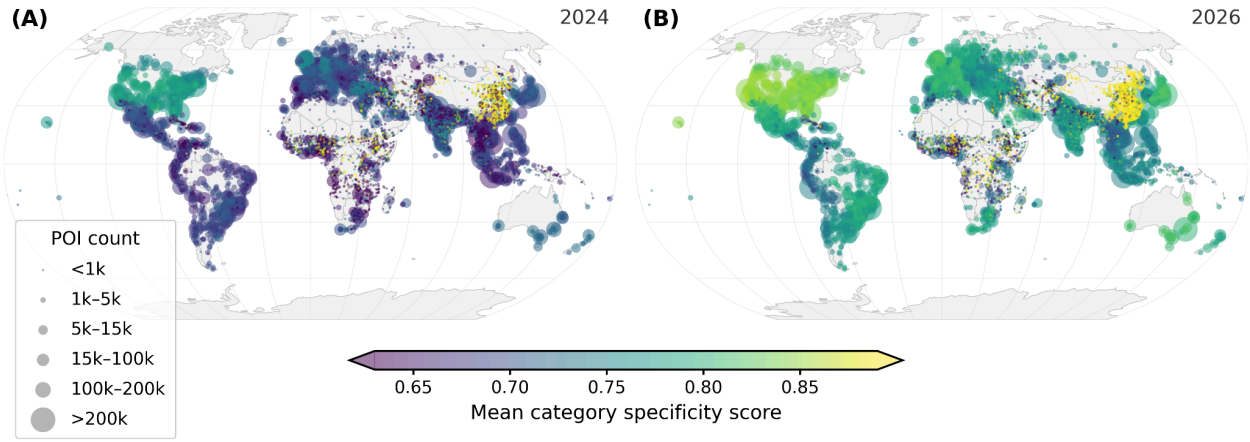


Fig. S1.2. City-level mean POI category specificity score (*S*□), November 2024 (A) and April 2026 (B). Each dot represents one GHSL urban area (*n* = 11,422); dot area is proportional to the total number of Overture Places POIs in the city. Color encodes the city-level mean specificity score S□ ∈ [0, 1], where higher values indicate POIs assigned to more granular taxonomy nodes (see Supplementary Methods §3.2 for the S□ definition). The color scale is clipped to the 10th - 90th percentile range (0.63–0.89) to maximize contrast. The shift from 2024 (A) to 2026 (B) reflects a near-universal improvement in category specificity: the global city-level median rose from S□ = 0.69 (2024) to S□ = 0.78 (2026). Gains were largest in Eastern and South-Eastern Asia (+0.15) and smallest in Northern America (+0.07). Residual low-specificity cities in 2026 are concentrated in Latin America and the Caribbean and Sub-Saharan Africa.

Table S1.3. Overture Buildings intrinsic quality by world region, 2024 and 2026. Regional summaries computed across all 11,422 GHSL-defined urban areas grouped into nine world regions. Numerical cells report the city-level median with the city-level mean in parentheses; Δ columns report the change in city-level median between the November 2024 and April 2026 data releases. n Buildings: total building feature count within GHSL urban areas. B-01: Dominant Source: the source contributing the largest regional building count, with that share in brackets. B-01: # Sources: number of unique data sources contributing buildings per city. B-01: Evenness: Pielou's evenness index E ∈ [0, 1]. B-02: % Subtype, % Height: city-level percentage of buildings carrying a recorded subtype or height attribute. B-03: % Overlap: city-level percentage of building footprints that geometrically intersect at least one other building footprint (area-based; see Supplementary Methods 5). EAB = East Asian Buildings

				B-01 (Data Sources)								B-02 (Attribute Completeness)						B-03 (Overlap)		
	Number of Buildings			Number of Sources			Dominant Source		Evenness			% Subtype			% Height			% Overlap		
Region	2024	2026	Δ	2024	2026	Δ	2024	2026	2024	2026	Δ	2024	2026	Δ	2024	2026	Δ	2024	2026	Δ
Australia and New Zealand (ANZ)	5,371,082	5,475,479	+1.9%	2.0 (2.1)	2.0 (2.1)	-	Microsoft ML Buildings [55.4%]	Microsoft ML Buildings [51.2%]	0.44 (0.51)	0.47 (0.54)	+0.03 (+0.03)	9.6 (14.3)	11.0 (16.7)	+1.4 (+2.4)	74.8 (59.3)	69.2 (57.2)	-5.6 (-2.1)	1.1 (1.5)	0.6 (1.3)	-0.5 (-0.2)
Central and Southern Asia (CSA)	117,653,751	111,804,834	-5.0%	3.0 (2.6)	3.0 (2.6)	-	Google Open Buildings [60.1%]	Google Open Buildings [57.7%]	0.54 (0.48)	0.56 (0.49)	+0.02 (+0.01)	0.0 (2.0)	0.0 (2.0)	-	0.0 (0.4)	0.0 (0.4)	-	19.3 (19.7)	16.6 (17.2)	-2.7 (-2.5)
Eastern and South-Eastern Asia (ESEA)	95,451,480	137,012,934	+43.5 %	1.0 (1.6)	2.0 (2.2)	+1 (+0.6)	OSM [46.9%]	OSM [36.4%]	0.00 (0.21)	0.17 (0.33)	+0.17 (+0.12)	4.0 (19.5)	0.2 (2.6)	-3.8 (-16.9)	0.0 (0.9)	0.0 (0.3)	0 (-0.6)	0.0 (4.2)	0.8 (4.2)	+0.08 (0.0)
Europe	54,949,484	56,569,868	+2.9%	2.0 (1.9)	2.0 (2.0)	0 (+0.1)	OSM [78.0%]	OSM [80.3%]	0.65 (0.60)	0.61 (0.58)	-0.04 (-0.02)	25.6 (30.9)	28.6 (33.1)	+3.0 (+2.2)	1.0 (19.6)	1.3 (16.5)	+0.3 (-3.1)	0.7 (1.6)	0.6 (1.3)	-0.1 (-0.3)
Latin America and the Caribbean (LAC)	103,907,810	103,574,010	-0.3%	3.0 (2.9)	3.0 (2.9)	-	Google Open Buildings [58.2%]	Google Open Buildings [55.2%]	0.62 (0.59)	0.63 (0.60)	+0.01 (+0.01)	0.4 (2.2)	0.5 (2.6)	+0.1 (+0.4)	0.0 (0.7)	0.0 (0.8)	0 (+0.1)	21.4 (23.1)	20.2 (21.7)	-1.2 (-1.4)
Northern Africa and Western Asia (NAWA)	33,896,053	33,179,668	-2.1%	2.0 (2.2)	2.0 (2.2)	-	Microsoft ML Buildings [50.3%]	Microsoft ML Buildings [51.8%]	0.52 (0.46)	0.57 (0.50)	+0.05 (+0.04)	0.3 (4.0)	0.4 (4.3)	+0.1 (+0.3)	0.0 (0.3)	0.0 (0.5)	0 (+0.2)	9.2 (10.5)	8.4 (9.1)	-0.8 (-1.4)
Northern America (NA)	43,964,116	44,440,854	+1.1%	2.0 (2.3)	2.0 (2.3)	-	OSM [70.9%]	OSM [77.3%]	0.50 (0.52)	0.46 (0.49)	-0.04 (-0.03)	11.2 (22.6)	15.2 (25.6)	+4.0 (+3.0)	82.6 (77.5)	80.1 (75.4)	-2.5 (-2.1)	0.2 (0.5)	0.1 (0.4)	-0.1 (-0.1)
Oceania	174,807	184,248	+5.4%	1.0 (1.0)	1.0 (1.0)	-	OSM [100.0%]	OSM [100.0%]	0.00 (0.00)	0.00 (0.00)	-	13.1 (20.9)	18.7 (23.9)	+5.6 (+3.0)	0.0 (11.1)	0.0 (11.0)	0 (-0.1)	0.1 (3.3)	0.2 (3.5)	+0.1 (+0.2)
Sub-Saharan Africa (SSA)	95,019,017	92,266,893	-2.9%	3.0 (2.8)	3.0 (2.8)	-	Google Open Buildings [54.4%]	Google Open Buildings [48.8%]	0.57 (0.54)	0.58 (0.56)	+0.01 (+0.02)	0.0 (1.2)	0.1 (1.2)	+0.1 (0)	0.0 (0.0)	0.0 (0.0)	-	10.8 (13.4)	9.0 (11.8)	-1.8 (-1.6)
Global	550,387,600	584,508,788	+6.2%	2.0 (2.3)	2.0 (2.4)	0 (+0.1)	Google Open Buildings [39.6%]	OSM [36.4%]	0.51 (0.44)	0.53 (0.48)	+0.02 (+0.04)	0.4 (10.5)	0.2 (6.4)	-0.2 (-3.9)	0.0 (5.5)	0.0 (4.8)	0 (-0.7)	7.5 (11.4)	6.0 (10.3)	-1.5 (-1.1)

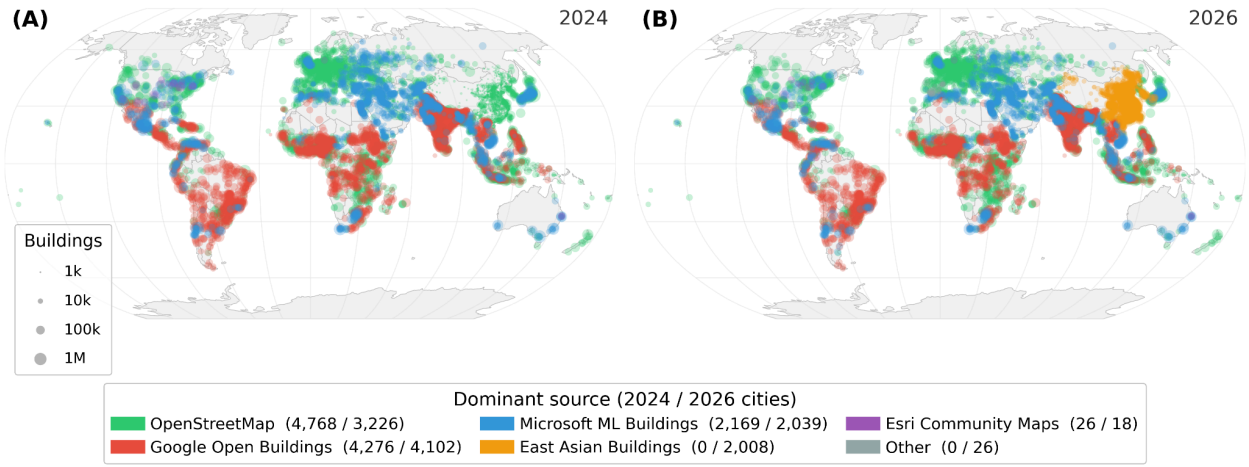


Fig. S1.3. Dominant building data source by city, November 2024 (A) and April 2026 (B). Each dot represents one GHSL urban area; dot size is proportional to city-level building count (scale: 1k–1M buildings). Color indicates the source contributing the majority of buildings to the city. Source counts are given in the legend for both data releases. The most visible change between releases is the emergence of the East Asian Buildings dataset (EAB) as the dominant source across Eastern and South-Eastern Asia in 2026, displacing Microsoft ML Buildings (blue), which had been the principal source for that region in 2024. Source geography elsewhere (i.e. OpenStreetMap leading in Europe and Northern America; Google Open Buildings leading in Sub-Saharan Africa and Latin America) remained stable between releases.

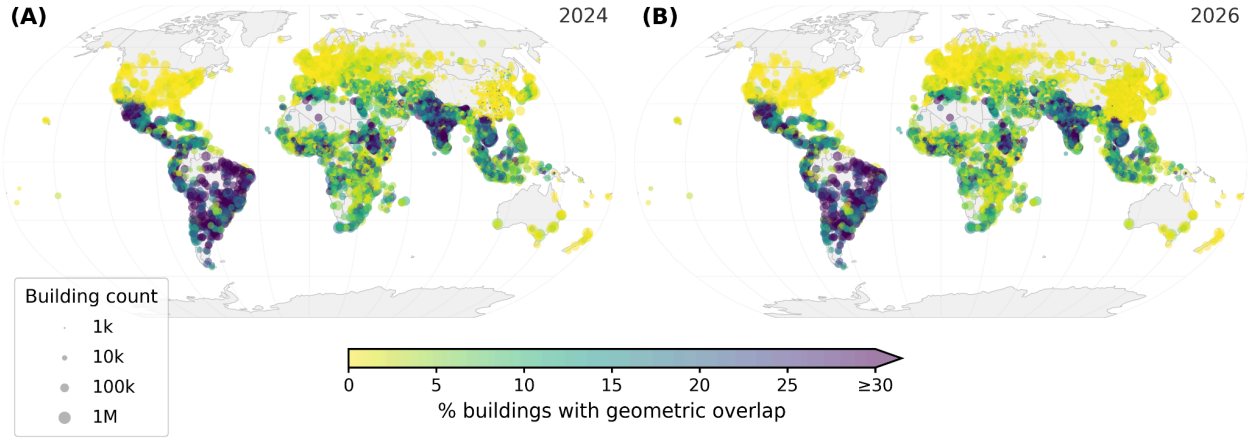


Fig. S1.4. Building geometric overlap by city, November 2024 (A) and April 2026 (B). Each dot represents one GHSL urban area; dot size is proportional to city-level building count (scale: 1k–1M). Color indicates the percentage of buildings whose footprint intersects at least one other building footprint by non-zero area (B-03 metric; see Supplementary Methods 4). The color scale is capped at 30% (global p95 = 32.3% in 2026); values exceeding the cap are shown at the upper limit. The geographic pattern is structurally stable across releases, suggesting that geometric overlap reflects source composition rather than release-specific processing. Latin America and the Caribbean and Central and Southern Asia (where Google Open Buildings and Microsoft ML Buildings dominate) exhibit systematically high overlap (medians 20.2% and 16.6% respectively in 2026), consistent with known footprint-bleed artefacts of satellite-imagery-derived ML detectors. Europe, Northern America, and Eastern and South-Eastern Asia (dominated by OpenStreetMap and the East Asian Buildings dataset) show near-zero overlap (medians <1%). The slight lightening of Eastern and South-Eastern Asia in 2026 reflects partial substitution of Microsoft ML Buildings by the East Asian Buildings dataset, whose footprints carry marginally less geometric overlap.

Table S1.4 provides distribution-free effect sizes for all six intrinsic quality metrics using Cliff's delta (δ), which is the proportion of cities that improved between releases minus the proportion that worsened, ranging from −1 (all cities deteriorated) to +1 (all cities improved)[1]. Cliff's delta makes no assumption about the shape of the underlying distribution and is therefore appropriate for the right-skewed, zero-bounded city-level metrics reported in this study. Source evenness (P-01) is excluded from this comparison because the number of contributing sources (*S*) changed structurally between releases, making release-to-release differences in $E = H/\ln(S)$ a result of the changed denominator rather than a quality change. All three POI quality sub-metrics showed large effects (δ = 0.73–0.85), as did building geometric overlap and subtype completeness (δ = 0.42–0.43), exceeding the conventional large-effect threshold of δ ≥ 0.42[2].

Table S1.4. Distribution-free effect sizes for release-to-release change in intrinsic quality metrics, computed using Cliff's delta (δ) across all cities present in both the November 2024 and April 2026 releases. δ = *(number of cities that improved − number that worsened) / n*, where improvement denotes a decrease for metrics in which lower values indicate higher quality (P-02 co-location, P-03 % Null category and B-03 % geometric overlap) and an increase for metrics in which higher values indicate higher quality (P-03 Specificity, B-02 % Subtype and % Height). Cities with no change between releases contribute to neither count. Effect size classification follows Vargha & Delaney (2000)[2]: negligible δ < 0.12, small 0.12 ≤ δ < 0.28, medium 0.28 ≤ δ < 0.42, large δ ≥ 0.42. n: number of cities with non-missing values in both data releases.

Metric	δ	Effect	% improved	% worsened	n
P-02 (% co-located)	0.73	Large	81%	8%	9,982
P-03 (% Null category)	0.85	Large	89%	4%	9,982
P-03 (Specificity)	0.79	Large	89%	10%	9,883
B-02 (% Subtype)	0.42	Large	63%	21%	11,239
B-02 (% Height)	0.06	Negligible	21%	16%	11,239
B-03 (% geometric overlap)	0.43	Large	68%	26%	11,422

Table S1.5. Intrinsic quality breakdown of Overture Places computed across GHSL-defined urban areas (*n*=11,422), for the "Other" category reported in main Table 1. Values report city-level median (mean).

				P-02 (Co-location)		P-03 (Category Quality)	
Source	Year	Number of POIs	% Share	Max Stack	% Co-located	% Null Cat.	Specificity (S□)
AllThePlaces	2026	828,832	2.0	1.0 (1.7)	0.0 (6.1)	0.0 (0.0)	0.90 (0.90)
DAC	2026	76,296	0.2	1.0 (1.4)	0.0 (6.2)	0.0 (9.4)	0.90 (0.88)
PinMeTo	2026	59,234	0.1	1.0 (1.1)	0.0 (0.7)	0.0 (16.3)	0.93 (0.91)
RenderSEO	2026	1,977	<0.1	1.0 (1.3)	0.0 (9.0)	0.0 (3.3)	0.92 (0.91)
Krick	2026	247	<0.1	1.0 (1.2)	0.0 (3.2)	0.0 (0.0)	0.92 (0.91)

Table S1.6. Intrinsic quality breakdown of Overture Buildings by data source. Minor sources omitted from main Table 2 are shown here.

Source	Year	Buildings (n)	% Share	% Subtype	% Height	% Num Floors
Instituto Geográfico Nacional (España)	2026	802,744	0.1	6.0	9.1	82.4
City of Vancouver	2026	16,587	0.0	0.0	100	0.0

2. Data download and indexing pipeline

Analyses were implemented in Python 3.12 using DuckDB 1.5.1 [3] with its spatial and H3 community extensions for in-memory columnar spatial analytics. GLMM models (Supplementary Material §6) were built using R (v4.6.0) and the glmmTMB R package [4]. Building and place data were stored and processed as GeoParquet files (zstd compression). City-level H3[5] assignments used h3-python v4.4.2; ISEA4H equal-area grid [6,7] operations used the vgrid C bridge via dggrid4py[8] as implemented in dggs-bench[9,10]. All scripts are available on GitHub at https://github.com/gatorlab-geo/overture-qa and on Zenodo at https://doi.org/10.5281/zenodo.22114268. Fig. S2.1. Illustrates the data processing and analysis pipeline.

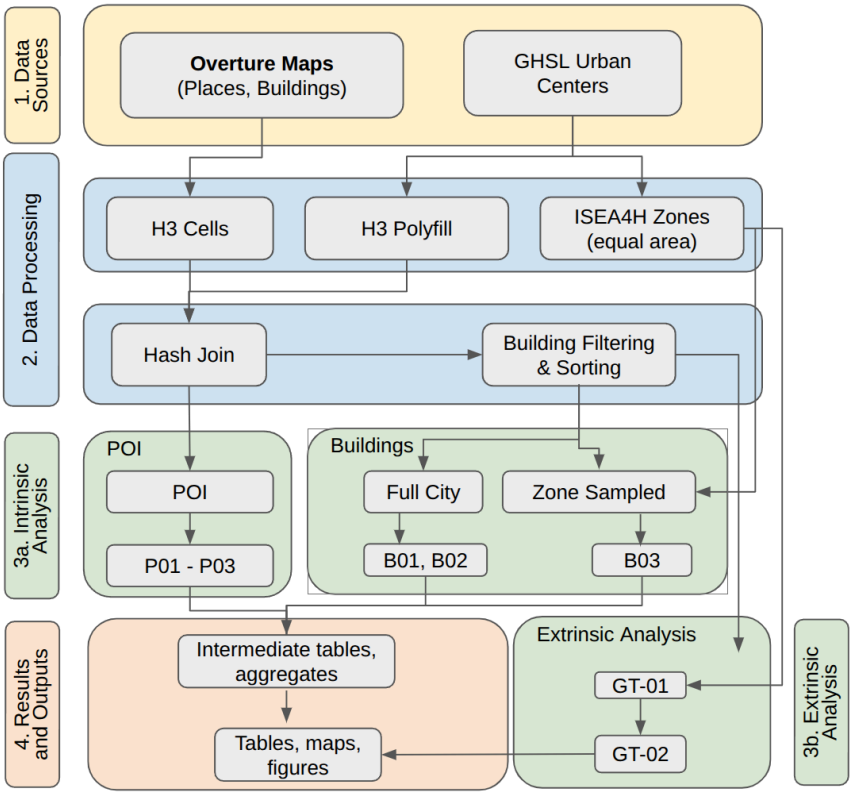


Fig. S2.1. Schematic diagram of the data processing and analysis pipeline

2.1. Data download

The two Overture releases analysed in this study were retrieved using different backends due to Overture's 60-day data retention policy, which removes older releases from the foundation's public Amazon S3 bucket. The April 2026 release (2026-04-15.0, schema v1.16.0) was downloaded directly from the Overture public S3 bucket (`s3://overturemaps-us-west-2/release/2026-04-15.0/`) using the AWS Command Line Interface (`aws s3 sync`), which supports resumable, incremental transfers. The November 2024 release (2024-11-13.0, schema v1.3.0) was no longer available on Overture's S3 at the time of analysis and was instead retrieved from the Source Cooperative mirror

(`https://data.source.coop/fused/overture/2024-11-13-0/`), a permanent archiving service that redistributes Overture releases without expiry. Both the Places and Buildings themes were downloaded for each release. Approximate raw data volumes were 20 GB (Places 2026), 18 GB (Places 2024), 118 GB (Buildings 2026), and 100 GB (Buildings 2024). All files are in GeoParquet format with zstd compression.

The Source Cooperative mirror uses a flat sequential file naming convention (`0.parquet`, `1.parquet`, …) that differs from the hive-partitioned layout on Overture's S3. Buildings data on Source Cooperative are additionally sub-partitioned into five directories (`part=0/` through `part=4/`). Downloads were implemented as resumable processes: the S3 backend skips files already present on disk, and the Source Cooperative backend writes each file atomically via a `.tmp` staging step where the target file becomes visible only upon successful completion, so interrupted transfers do not leave partially written output. The Source Cooperative files also carry an additional `part` metadata column absent from the S3 schema; this column was dropped during download to maintain schema compatibility between releases.

2.2. H3 city assignment index

To assign each place and building feature to a GHSL urban settlement [11] without repeated polygon-containment joins, a one-time H3 resolution-9 polyfill was computed for all 11,422 GHSL urban settlement polygons using `scripts/00b_index_h3.py`. At resolution 9, each H3 hexagonal cell covers approximately 0.105 km², and cell edge length is approximately 174 m . This is well below the footprint of even the smallest GHSL urban centres. For each city, H3 cells were identified whose centroids fall within the GHSL polygon boundary (centroid containment is the H3 standard; boundary error < 0.1% at city-level aggregation). For MultiPolygon geometries, each part was polyfilled separately and the results unioned. The resulting lookup table (h3_cell → city_id, region, stratum) was stored as a zstd-compressed Parquet file (`data/dggs/ghsl_h3_res9.parquet`). This index replaces ST_Within spatial joins in all downstream analysis scripts (P-01 to P-03, B-01 to B-03) with pure hash joins on the `h3_cell` column, substantially reducing query runtime at the DuckDB layer [12]. A duplicate-free guarantee is enforced at index construction time: boundary-overlap cells shared by adjacent cities are assigned to the first city encountered during iteration, and duplicates are dropped. The left panel in Fig. S2.2 illustrates the H3 polyfill results for Budapest, Hungary. Regions were assigned to cities using the `GC_DEV_USR_2025` field of the GHSL layer, which partitions cities into nine SDG-aligned world regions (see main Methods and Supplementary Materials §5 for details on the nine-region classification).

2.3. ISEA4H analysis zone index

For the building overlap analysis (B-03) and the extrinsic ground-truth assessment, a separate spatial index was required that provides a fixed, latitude-invariant analysis zone centred on each city's urban core, independent of the city polygon's shape or extent. This was computed using the Icosahedral Snyder Equal Area Aperture 4 Hexagonal (ISEA4H) discrete global grid system

at resolution 14 (cell area: 0.19 km²), implemented via a compiled C bridge to the DGGRID library as described in [10].

For each of the 11,422 GHSL cities, the anchor point was taken as the Shapely `representative_point()` (equivalent to PostGIS `ST_PointOnSurface`), which is guaranteed to lie within the polygon boundary. This avoids misplaced analysis zones for coastal or concave-shaped cities where the geometric centroid may fall outside the polygon. The anchor point was encoded to its containing ISEA4H cell at resolution 14. Because ISEA4H is an equal-area projection, this single cell (~0.19 km²) covers an identical physical area for every city on Earth, regardless of latitude. The resulting index (city_id → ISEA4H cell identifier, anchor coordinates) was stored as a Parquet file (`data/dggs/all_cities_isea4h_k0.parquet`).
Fig. S2.2. illustrates the ISEA4H indexing scheme for Budapest, Hungary.

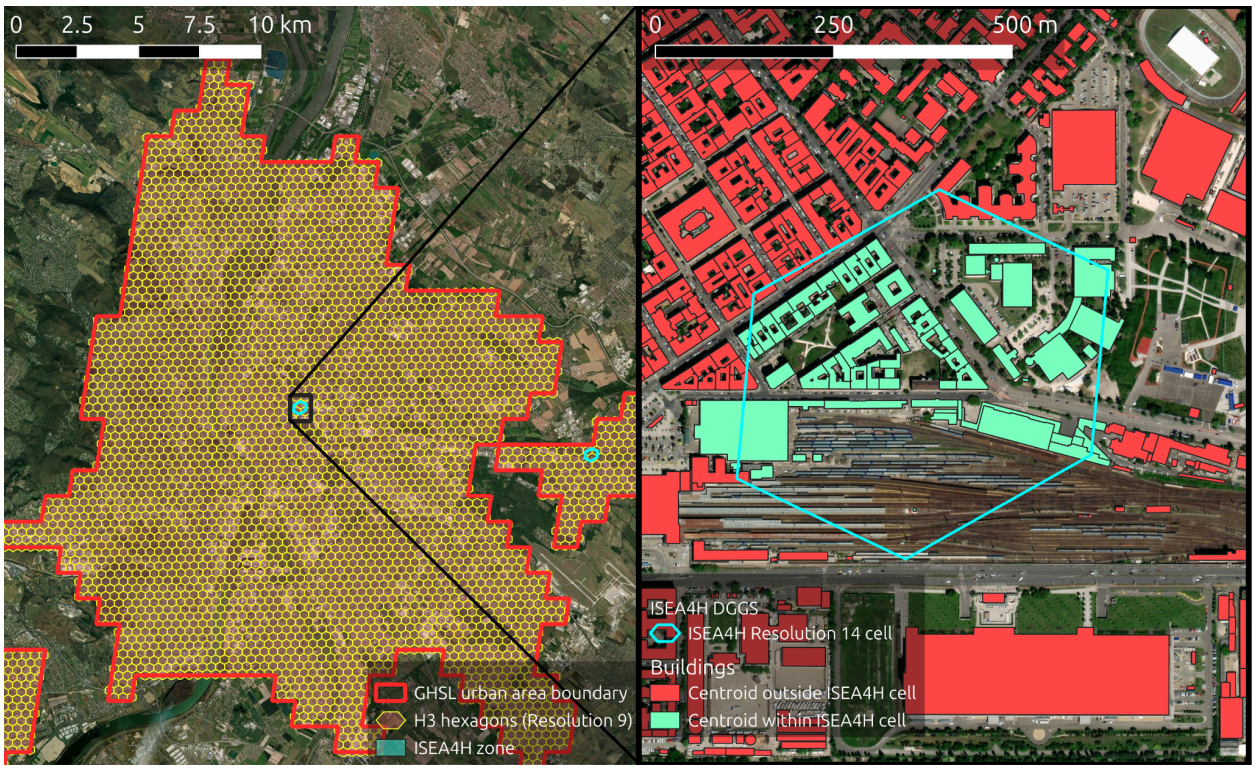


Fig. S2.2: DGGS indexing layers for Budapest, Hungary. **Left:** GHSL urban area boundary (red outline), H3 Resolution-9 hexagons used for city-level feature assignment (yellow cells), and the ISEA4H analysis zone (teal hexagon) centred on the urban core. **Right**: zoomed inset centered on the ISEA4H Resolution-14 cell (cyan boundary). Buildings whose centroid falls within the cell boundary are highlighted in teal; buildings whose centroid falls outside are shown in red. Buildings in teal were included in the manual ground-truth completeness and precision assessment for this city. The ISEA4H Resolution-14 cell (~0.19 km²) provides an identical physical sampling area for every city on Earth regardless of latitude, ensuring unbiased comparability across regions.

3. Places taxonomy and specificity score

3.1. Taxonomy structure

Overture Maps Places features are classified using a hierarchical taxonomy organized into 22 top-level (Level 1) category branches, which are further subdivided into 2,095 named subcategories spanning Levels 2 through 6. The taxonomy is publicly versioned in the Overture schema repository [13]. Level 1 branches vary substantially in breadth and depth. For example, `retail` contains 286 subcategories across 5 levels, whereas `accommodation` contains only 17 subcategories across 3 levels, and `private_establishments_and_corporates` contains 4 subcategories that do not extend beyond Level 2 (Table S3.1).

Table S3.1. Number of sub-categories falling under 22 Level 1 place categories

Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	Total
retail	21	133	111	21	0	286
professional_services	155	95	19	0	0	269
business_to_business	11	131	81	12	2	237
eat_and_drink	3	150	73	0	0	226
health_and_medical	89	85	17	0	0	191
active_life	4	103	18	0	0	125
attractions_and_activities	68	18	12	0	0	98
home_service	78	14	0	0	0	92
arts_and_entertainment	52	32	0	0	0	84
travel	17	60	3	0	0	80
automotive	18	52	3	0	0	73
education	16	47	3	0	0	66
public_service_and_government	38	15	12	0	0	65
beauty_and_spa	26	14	0	0	0	40
financial_service	21	11	0	0	0	32
real_estate	21	6	0	0	0	27
pets	6	17	3	0	0	26
mass_media	6	16	0	0	0	22
religious_organization	11	7	0	0	0	18
accommodation	16	1	0	0	0	17
structure_and_geography	17	0	0	0	0	17
private_establishments_and_corporates	4	0	0	0	0	4
Total	**698**	**1007**	**355**	**33**	**2**	**2095**

To assess whether the taxonomy changed between the two data releases, we examined the full commit history of `overture_categories.csv` in the Overture schema repository. A review of the commit history for that category taxonomy file confirmed that no structural changes (branch additions, removals, reclassifications, or depth changes) were made to the taxonomy during the period spanning both releases. The 22 Level 1 branches, 2,095 subcategories, and all branch depths were identical across releases, supporting a direct cross-release comparison of S_k.

3.2. Specificity score (S_k)

Overture's category taxonomy depth (the integer level 1–6 of an assigned category) provides an intuitive proxy for how specifically a place has been categorized. However, raw depth is not comparable across Level 1 branches because branches differ substantially in their total subcategory count and maximum depth. A Level 2 assignment (e.g., `restaurant` within `eat_and_drink`) represents a coarse label within a branch of 226 subcategories spanning four levels, while a Level 2 assignment (e.g., `hotel` within `accommodation`) represents a relatively specific label within a branch of only 17 subcategories spanning three levels. Treating both as equivalent "depth-2" labels ignores categorization specificity with branch architecture.

To address this, we adopted a branch-aware log-normalized specificity score S□, which measures how far down a given L1 branch an assigned category label is, relative to the full depth of that branch. For a place assigned to a category at level k within a branch rooted at Level 1 category b, S□ is defined as:

$$S□ = log(1 + \Sigma□_{=2}{}^{k} n□) / log(1 + \Sigma□_{=2}{}^{L} n□)$$

Where n_l is the number of taxonomy categories at level *l* within branch *b*, and *L* is the maximum depth of that branch and log denotes the natural logarithm. The score equals zero when $k = 1$ (only the Level 1 label is known) and equals one when $k = L$ (the most specific available label within the branch is assigned). The logarithm compresses the contribution of each additional level, reflecting diminishing informational returns as specificity increases.

As a worked example, consider a place assigned to `restaurant` (Level 2, `eat_and_drink` branch) versus a place assigned to `hotel` (Level 2, `accommodation` branch). For `eat_and_drink`, $n_2 = 3$, $n_3 = 150$, $n_4 = 73$, so:

$$S_2^{eat_and_drink} = log(1 + 3) / log(1 + 3 + 150 + 73) = log(4) / log(227) \approx 0.256$$

For `accommodation`, $n_2 = 16$, $n_3 = 1$, so

$$S_2^{accommodation} = log(1 + 16) / log(1 + 16 + 1) = log(17) / log(18) \approx 0.980$$

A `hotel` assignment is therefore scored as nearly maximally specific within its branch (0.980), while a `restaurant` assignment (equally deep in the raw taxonomy) is scored as low-specificity (0.256), correctly reflecting that `eat_and_drink` offers 223 more-specific subcategories that have not been utilized.

S□ scores were pre-computed for all 2,095 taxonomy nodes and stored. At analysis time, each Place feature in the 2024 and 2026 releases was joined to the taxonomy table. Places with a null primary category, or with a category string not present in the taxonomy, were excluded from the S□ mean and tallied separately as `pct_null_category` and `pct_unmatched_category`,

respectively. The city-level median and mean specificity scores for both releases, broken down by region are shown in Table S1.2., while Fig. S1.2. maps the spatial distribution of city-level S_k.

4. Building overlap computation (B-03): Batched city-loop architecture

4.1. Motivation

Building geometric overlap (B-03) measures the percentage of building footprints within a city whose polygon boundary intersects at least one other building's boundary with a positive-area intersection (i.e., true 2D overlap, not shared walls or corner touches). Quantifying this at scale across 11,422 urban centres for two releases requires evaluating spatial self-joins. This operation's complexity scales quadratically with the number of buildings per city.

A naïve single-pass approach, i.e. filtering buildings by region and performing a `ST_Intersects` self-join across all buildings in that region was attempted during development and failed at production scale. Region-level datasets contain from 33 million (Northern Africa and Western Asia) to 137 million (Eastern and South-Eastern Asia) building footprints in full urban extents, or 1 million to 12 million when restricted to ISEA4H analysis zones. Both approaches were attempted. DuckDB builds a transient STR-tree spatial index on the probe side of a spatial join, but at either scale the memory footprint of the index and the on-disk spill for the hash-join driver exhausted all available NVMe storage even with a 9-region partition strategy.

4.2. City-loop architecture

The solution adopted is a city-loop with per-city spatial containment: rather than filtering by region, each of the 11,422 cities is processed independently. For each city, buildings are further restricted to a fixed-area zone (see Supplementary Material §2.3) centered on an inside representative point, yielding a spatially bounded subset of manageable size. For the building overlap computation, a fixed k=2 zone (19 cells, ~3.61 km²) is applied uniformly to all 11,422 cities. Unlike the manual counting protocol, where zone size was constrained by the effort of visual inspection, the automated overlap computation scales freely, and the larger zone provides a more statistically stable sample of building pairs per city.

For each city, the computation proceeds in two steps:

1. **Zone materialization:** All building footprints whose geometries intersect the city's ISEA4H k=2 zone polygon (19 cells, ~3.61 km²) are loaded into a temporary in-memory table, capped at 15,000 buildings per city to bound GEOS memory allocation. This cap affected only 0.5% (55 out of 11,422) cities.

2. **Area-discriminated overlap**: A self-join on the temporary table evaluates all building pairs where `ST_Intersects` is true and `id_a < id_b` (upper-triangle only, to avoid double-counting). For each candidate pair, `ST_Area(ST_Intersection(a.geometry, b.geometry))` is computed. Only pairs with intersection area > 0 are retained as true overlaps (excluding shared boundary lines and corner touches). The metric reported is the fraction of distinct buildings involved in at least one true overlap within the zone.

4.3. Memory management, resumability, and batched execution

DuckDB delegates all spatial predicate and polygon operations to the GEOS C library. GEOS allocates STR-tree index structures and prepared-geometry caches in native (OS) memory outside DuckDB's buffer manager and the `memory_limit` session setting has no effect on these allocations. Over thousands of sequential city iterations, GEOS memory grows monotonically and does not release between queries. This behavior was confirmed empirically on the analysis machine (58 GB RAM), causing an out-of-memory crash at approximately 2,000 cities when processed in a single continuous run.

Two complementary mechanisms address this. First, within-process connection recycling, i.e. every 10 cities, the active DuckDB connection is explicitly closed, Python garbage collection is triggered (`gc.collect()`), and a fresh connection is opened. Destroying the connection releases DuckDB's buffer pool and tears down the GEOS library context, allowing the operating system to reclaim all native allocations. Second, process-level batching, that is the shell wrapper `run_b03_batched.sh` re-launches the Python interpreter every 10 cities (`--max-cities 10`), ensuring a complete process exit and full memory release between batches regardless of any residual GEOS state. Together these two layers form a memory management strategy that allows processing the entire dataset on a consumer machine.

Each city's result is written to an individual checkpoint file at `data/intermediate/b03_cache/{year}/{city_id}.parquet` immediately upon completion. On any restart (whether from a crash, OOM kill, or deliberate interruption), cities whose checkpoint file already exists are skipped, so the analysis resumes exactly where it stopped with no data loss. Final aggregation of per-city checkpoint files into regional means is performed in a single pass by `validate_b03.py --aggregate` after all cities are complete.

5. Ground-truth sampling design and adaptive zone protocol

5.1. Stratification framework

The ground-truth building completeness assessment requires cities that are representative of global variation in Overture data quality. This variation is driven both by geographic region and

by city size. We used a two-dimensional stratification: (1) the nine macro-geographical regions as implemented in the GHSL 2025 Urban Center database (Fig. S5.1), and (2) the four GHSL functional urban area size classes: Large Metropolitan Area, Metropolitan Area, Medium Urban Area, and Small Urban Area.

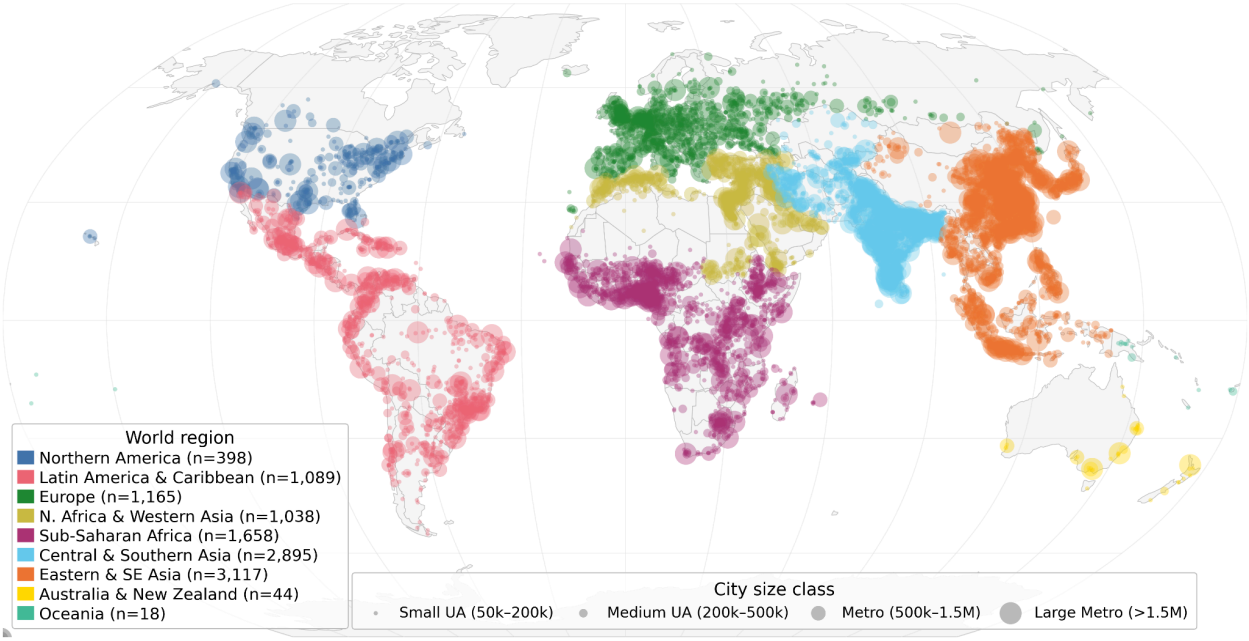


Fig. S5.1. Geographic extent and regional classification of the study area. Country fill colours indicate the nine macro-geographical region assigned to each country via the GC_DEV_USR_2025 field of the GHSL 2025 Urban Centre Database. White dots mark the 11,422 GHSL urban centres included in the analysis. Countries shown in grey contain no GHSL urban area and are excluded from all analyses. The nine-region classification is a modified version of the United Nations SDG super-regional framework in which Europe and Northern America form a single group. Region city counts: Eastern and South-Eastern Asia (n = 3,117), Central and Southern Asia (2,895), Sub-Saharan Africa (1,658), Europe (1,165), Latin America and the Caribbean (1,089), Northern Africa and Western Asia (1,038), Northern America (398), Australia and New Zealand (44), Oceania (18).

Crossing 9 regions by 4 size classes yields a maximum of 36 strata. However, Oceania has no cities classified as Large Metropolitan or Metropolitan Area, reducing the realized strata to 34. Within each stratum, 10 random cities were selected for ground-truth evaluation. For the Australia–New Zealand and Oceania regions, all available cities were retained regardless of rank, to maximize representation in regions with limited Overture data coverage. The resulting ground-truth sample comprises 338 cities across 34 strata (Table S5.1). Of these, 18 Oceania cities are excluded from the binomial GLMM (Supplementary Materials §6) due to insufficient stratum sizes for regression; the GLMM therefore uses 320 cities.

Table S5.1: Ground-truth sample allocation by region and GHSL functional urban area size class. Cells show the number of cities selected (target = 10 per stratum; AUS/NZ and Oceania retain all available cities). Row totals are the count of validated cities per region.

Region	Large Metropolitan	Metropolitan	Medium Urban	Small Urban	Total
Australia and New Zealand	3	4	3	30	40
Central and Southern Asia	10	10	10	10	40

Eastern and South-Eastern Asia	10	10	10	10	40
Europe	10	10	10	10	40
Latin American and the Caribbean	10	10	10	10	40
Northern Africa and Western Asia	10	10	10	10	40
Northern America	10	10	10	10	40
Oceania	0	0	2	16	18
Sub-Saharan Africa	10	10	10	10	40
Total	73	74	75	116	338

5.2. Analysis zone

For each of the 338 sampled cities, Overture Buildings 2026 were exported for a fixed, equal-area ISEA4H cell centered on the GHSL representative point. The analysis zone was defined as the ISEA4H resolution-14 cell (~0.19 km²). This approach produced a median of 163 (mean: 226) mapped buildings per city.

Each city's export is a GeoPackage containing three layers: buildings_2026 (footprints clipped to the selected zone), zone (the dissolved ISEA4H boundary polygon), and places_2026 (Overture POI points within the zone).

5.3. Ground-truth building completeness and precision

For each exported zone, a trained analyst (one of the authors) visually inspected the Overture building footprints against satellite imagery to assess two quality dimensions: completeness (the proportion of real-world buildings captured by Overture) and precision (the proportion of Overture footprints that correspond to a real structure). The three quantities recorded per city are: n_b, the number of Overture building footprints within the zone; n_o, the number of real-world buildings visible in imagery that are absent from the Overture dataset (omissions); and n_c, the number of Overture footprints that do not correspond to any real structure (commissions). From these, true positives are defined as $TP = n_b - n_c$, false negatives as $FN = n_o$, and false positives as $FP = n_c$. Completeness and precision for city *i* are then:

$$Completeness = TP / (TP + FN) = (n_b - n_c)/(n_b - n_c + n_o)$$
$$Precision = TP / (TP + FP) = (n_b - n_c)/n_b$$

Fig. S5.2. illustrates this computation for a single zone (~0.19km^2) in Tirana, Albania, where $n_b = 183$, $n_o = 42$, and $n_c = 10$, resulting in a completeness of 0.805 and a precision of 0.945.

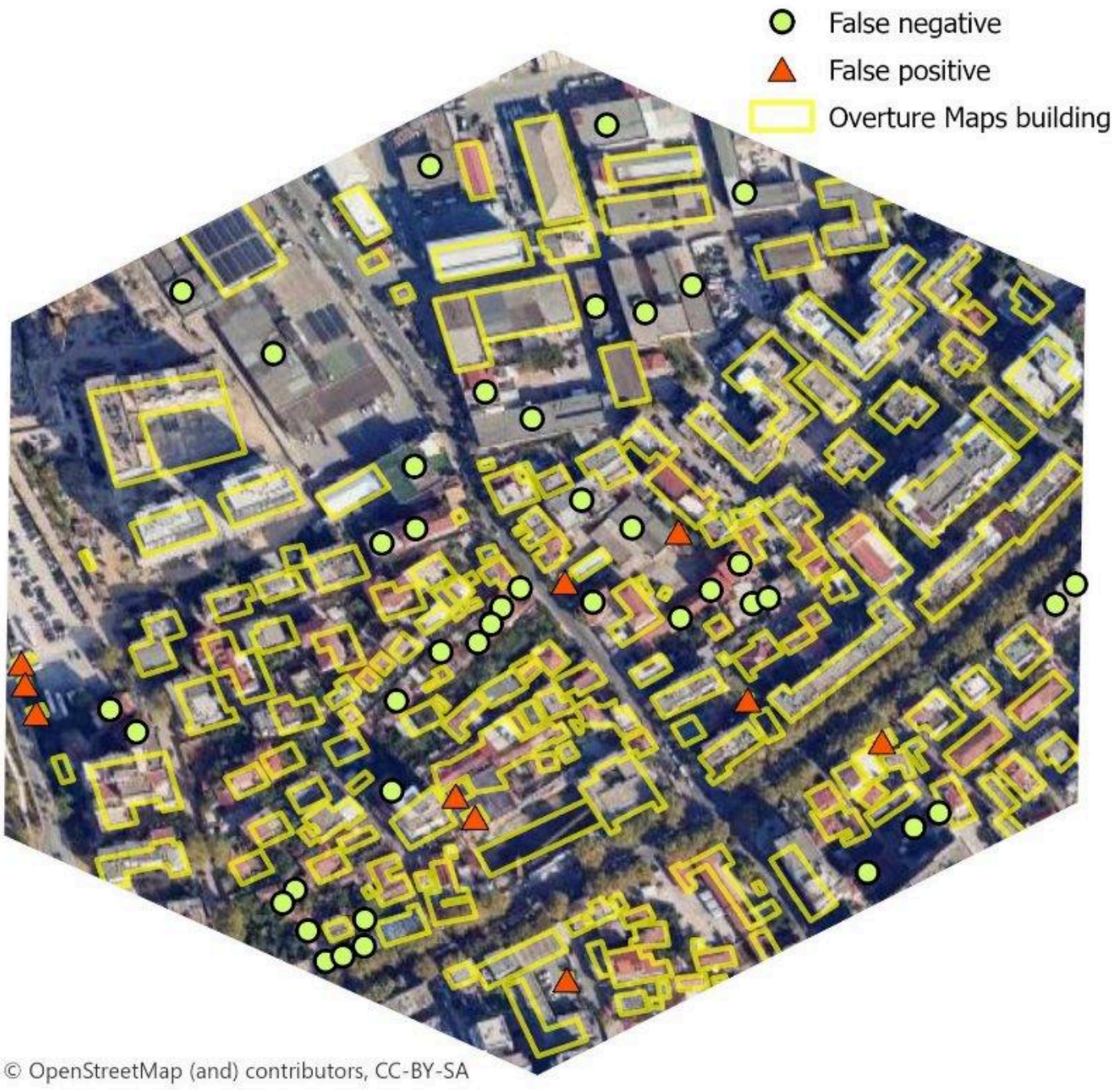


Fig. S5.2. Worked example of the ground-truth completeness and precision computation for a single ISEA4H analysis zone in Tirana, Albania (~0.19 km²). Yellow polygon outlines show Overture Maps building footprints ($n_b = 183$). Green circles mark false negatives ($n_o = 42$): real-world buildings visible in satellite imagery that are absent from the Overture dataset. Orange triangles mark false positives ($n_c = 10$): Overture features with no corresponding real-world structure. In this zone, completeness = 0.805 and precision = 0.945, indicating that Overture captured approximately 81% of real buildings while approximately 95% of its footprints correspond to actual structures.

6. Binomial GLMM

6.1. Model specification

To determine the effects associated with completeness and precision a binomial generalized linear mixed model (GLMM) was used. While the ratios for completeness and precision for a city introduced in Supplementary Materials §5.3 are useful for descriptive summaries and interpretation, they do not need to be computed explicitly for model fitting. Instead, the underlying counts of successes and failures were modeled directly using binomial GLMMs with a logit link, which preserves the numerator-denominator structure and properly accounts for variation in the number of buildings contributing to each city-level estimate. The model coefficients (β, etc.) are expressed in log-odds and can be exponentiated to obtain odds ratios, OR = e^{β}, which quantify the multiplicative change in the odds associated with a one-unit increase in a predictor. For completeness, the number of correctly mapped buildings in city i was modeled as

$$TP_i \sim Binomial(TP_i + FN_i, p_i^{(C)})$$

where $TP_i + FN_i$ is the total number of buildings in the reference dataset for city i, and $p_i^{(C)}$ is the model-based probability that a reference building is captured by the mapped dataset. For precision, the number of correctly mapped buildings was modeled as

$$TP_i \sim Binomial(TP_i + FP_i, p_i^{(P)})$$

where $TP_i + FP_i$ is the total number of buildings mapped in city i, and $p_i^{(P)}$ is the model-based probability that a mapped building is correct. In both models, the linear predictor included fixed effects for region, city size, standardized city population density and source composition, plus a random intercept for country:

$$logit\left(p_i^{(C)}\right) = \beta_0^{(C)} + \sum_{r=1}^{R-1} \gamma_r^{(C)} Region_{i,r} + \sum_{s=1}^{S-1} \delta_s^{(C)} CitySize_{i,s} + \mu^{(C)} Density_i + \propto_1^{(C)} lrOSM_i + \propto_2^{(C)} lrMS_i + \propto_3^{(C)} lrG_i + \propto_4^{(C)} lrZ_i + u_{country(i)}$$

$$logit\left(p_i^{(P)}\right) = \beta_0^{(P)} + \sum_{r=1}^{R-1} \gamma_r^{(P)} Region_{i,r} + \sum_{s=1}^{S-1} \delta_s^{(P)} CitySize_{i,s} + \mu^{(P)} Density_i + \propto_1^{(P)} lrOSM_i + \propto_2^{(P)} lrMS_i + \propto_3^{(P)} lrG_i + \propto_4^{(P)} lrZ_i + u_{country(i)}$$

where

- β_0 is the intercept
- γ_r are the fixed-effect coefficients for the region categories relative to the reference region
- δ_s are the fixed-effect coefficients for the city-size categories relative to the reference city-size class
- $\alpha_1 – \alpha_4$ are coefficients for the source-composition variables
- μ is the fixed-effect coefficient for standardized city population density

- $u_{country(i)} \sim N\left(0, \sigma_u^2\right)$ is the random intercept for the country to which city i belongs

Source composition was derived from five shares that sum to one within each city from the following sources: OpenStreetMap (osm), Microsoft (ms), Google (g), EAB (z), and an aggregated "other" (o) component combining ESRI Community Maps and Instituto Geográfico Nacional España (IGN). These shares are compositional data and therefore constrained to be nonnegative and to sum to one. Because of this constant-sum constraint, including the raw shares directly as linear predictors would induce perfect dependence. The shares were therefore transformed into four isometric log-ratio coordinates, which provide an unconstrained set of predictors for a five-part composition. Zero values were replaced by a small constant, ε = 10^{-6}, and the shares subsequently renormalized. The four coordinates were then computed as

$$lrOSM_i = \sqrt{\frac{5}{4}} log\left(\frac{osm}{(ms \bullet g \bullet z \bullet o)^{1/4}}\right)$$

$$lrMS_i = \sqrt{\frac{3}{4}} log\left(\frac{ms}{(g \bullet z \bullet o)^{1/3}}\right)$$

$$lrG_i = \sqrt{\frac{2}{3}} log\left(\frac{g}{(z \bullet o)^{1/2}}\right)$$

$$lrZ_i = \sqrt{\frac{1}{2}} log\left(\frac{z}{o}\right)$$

where osm, ms, g, z, and o, are the normalized, zero-adjusted source shares for city i. These log-ratio coordinates form an orthonormal representation of the five-part source composition, so only four predictors are needed. The omitted "other" term is therefore part of the reference structure of the compositional balance. Because these coefficients are expressed as log-ratio coordinates rather than raw proportions, they should be interpreted as relative shifts in source composition rather than independent marginal effects of single sources.

The logit link is defined as $logit(p) = log\left(\frac{p}{1-p}\right)$.

6.2. GLMM results

Table S6.1. Binomial mixed-effects model of building completeness. Base categories are Europe (Region) and small urban area (City size); Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05; highlighted in boldface.

Predictor	Estimate	SE	p
(Intercept)	3.524	0.357	**<0.001*** **
Population density (scaled)	0.120	0.019	**<0.001*** **
Region			
Australia and New Zealand	-0.835	0.913	0.361
Central and Southern Asia	-0.375	0.588	0.523
Eastern and South-Estern Asia	-0.929	0.532	0.081
Latin America and the Caribbean	1.601	0.570	**0.005*** **
Northern Africa and Western Asia	-0.881	0.503	0.08
Northern America	-0.203	0.907	0.823
Sub-Saharan Africa	1.806	0.544	**<0.001*** **
City Size			
Metropolitan Area	0.086	0.048	0.077

Large Metropolitan Area	-0.196	0.042	**<0.001*** **
Medium Urban Area	0.131	0.041	**0.002*** **
Source composition			
OSM log-ration coordinate	0.020	0.004	**<0.001*** **
Microsoft log-ration coordinate	-0.049	0.005	**<0.001*** **
Google log-ratio coordinate	-0.216	0.032	**<0.001*** **
EAB log-ration coordinate	-0.012	0.030	0.699

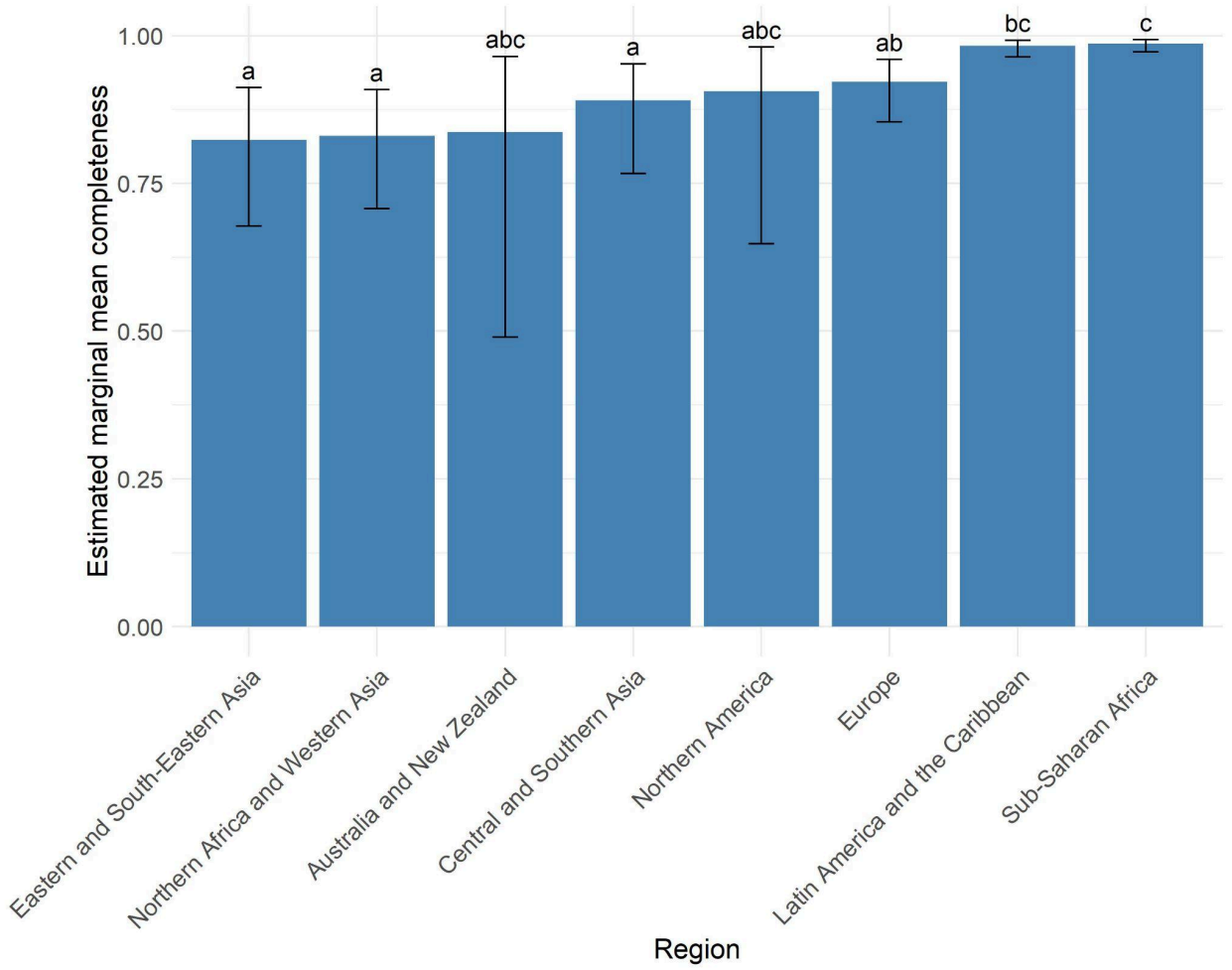


Fig. S6.1.Model-adjusted estimated marginal means of building completeness by world region, derived from the binomial mixed-effects model (Table S6.1). Estimates represent predicted completeness at the reference city size (small urban area) and mean population density and source composition, averaged across countries within each region. Error bars show 95% confidence intervals. Letters above bars indicate Tukey-adjusted pairwise groupings: regions sharing a letter are not significantly different from each other (α = 0.05).

Table S6.2. Binomial mixed-effects model of building precision. Base categories are Europe (Region) and small urban area (City size); Significance levels: *** p < 0.001, ** p < 0.01, * p < 0.05; highlighted in boldface.

Predictor	Estimate	SE	p
(Intercept)	3.358	0.271	**<0.001*** **
Population density (scaled)	-0.112	0.020	**<0.001*** **
Region			
Australia and New Zealand	1.058	0.693	0.127
Central and Southern Asia	-1.040	0.442	**0.019***
Eastern and South-Estern Asia	-1.647	0.408	**<0.001*** **
Latin America and the Caribbean	-0.300	0.429	0.485
Northern Africa and Western Asia	0.218	0.387	0.572
Northern America	0.875	0.690	0.205
Sub-Saharan Africa	-0.571	0.410	0.164
City Size			
Metropolitan Area	-0.134	0.056	**0.016***
Large Metropolitan Area	-0.260	0.052	**<0.001*** **
Medium Urban Area	-0.164	0.047	**<0.001*** **
Source composition			
OSM log-ratio coordinate	0.005	0.005	0.343
Microsoft log-ratio coordinate	-0.026	0.005	**<0.001*** **
Google log-ratio coordinate	-0.042	0.023	0.071
EAB log-ratio coordinate	0.153	0.023	**<0.001*** **

Below is a worked example to illustrate how regional and compositional effects combine in the precision model. Because EAB occurs in Eastern and South-Eastern Asia, the EAB log-ratio coordinate (β = 0.153, p < 0.001) partly offsets the otherwise lower precision for that region. For illustration, increasing the EAB share from 30% to 80% relative to the aggregated ESRI+IGN share would increase the model log-odds of precision by about 0.24, corresponding to roughly a 27% increase in the odds of precision, all else equal. Nonetheless, a city in Eastern and South-Eastern Asia with a 50% EAB share would still be expected to have lower precision overall than an otherwise similar European city, as the regional coefficient (β = −1.647) outweighs the positive EAB contribution of approximately +1.39 on the logit scale, leaving a net difference of roughly −0.26, corresponding to approximately a 23% decrease in the odds of precision.

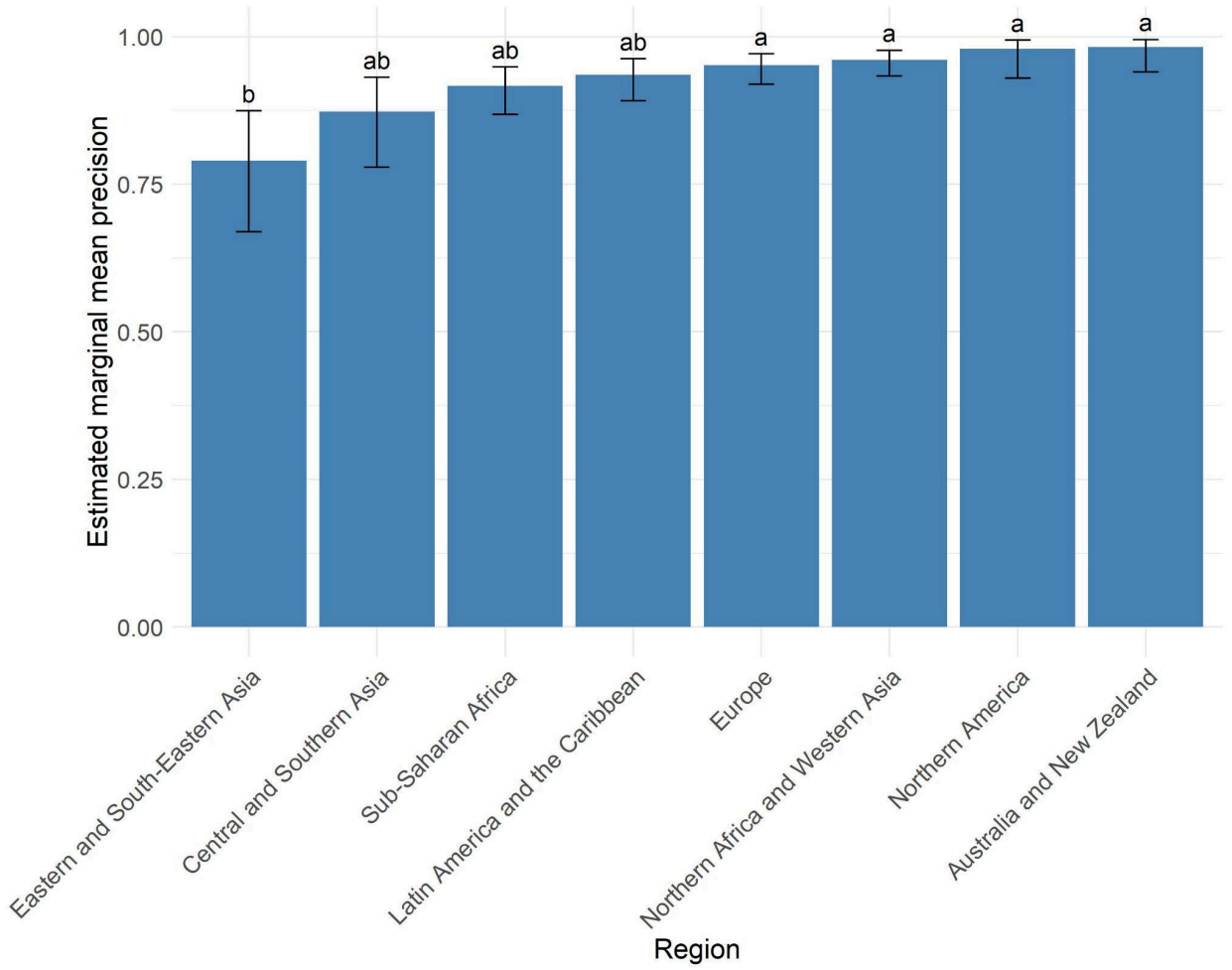


Fig. S6.2. **Model-adjusted estimated marginal means of building precision by world region, derived from the binomial mixed-effects model** (Supplementary Table 6.2). Estimates represent predicted precision at the reference city size (small urban area) and mean population density and source composition, averaged across countries within each region. Error bars show 95% confidence intervals. Letters above bars indicate Tukey-adjusted pairwise groupings: regions sharing a letter are not significantly different from each other (α = 0.05). Oceania is excluded from the model due to insufficient stratum sizes for regression.

6.3. Model validation

Grouped cross-validation by country was used to evaluate the predictive performance of the completeness and precision models under the same setting used for global prediction. All 320 cities were partitioned into 10 folds at the country level, such that all cities from a given country were assigned to the same fold. For each fold, the model was refit on the remaining countries and predictions for the held-out countries were generated using fixed effects only. This provides a more conservative assessment than ordinary cross-validation at the city level, because it evaluates transfer to entirely unseen countries rather than to additional cities from countries already represented in the training data.

Across the 10 folds, predictive accuracy was moderate. For completeness, the mean RMSE was 0.174 and the mean MAE was 0.112, corresponding to typical errors of about 17 and 11 percentage points on the 0–1 completeness scale. For precision, performance was better, with a mean RMSE of 0.117 and a mean MAE of 0.074, corresponding to about 12 and 7 percentage points on the 0–1 precision scale. These results indicate that the models retain useful predictive skill when applied to countries not included in model fitting.

References

1. Cliff, N. Dominance statistics: Ordinal analyses to answer ordinal questions. *Psychol. Bull.* **114**, 494–509 (1993).
2. Vargha, A. & Delaney, H. D. A Critique and Improvement of the CL Common Language Effect Size Statistics of McGraw and Wong. *J. Educ. Behav. Stat.* **25**, 101–132 (2000).
3. Raasveldt, M. & Mühleisen, H. DuckDB: an Embeddable Analytical Database. in *Proceedings of the 2019 International Conference on Management of Data* 1981–1984 (Association for Computing Machinery, New York, NY, USA, 2019). doi:10.1145/3299869.3320212.
4. Brooks, M., E. *et al.* glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *R J.* **9**, 378 (2017).
5. Brodsky, I. H3: Uber’s Hexagonal Hierarchical Spatial Index. *Uber Blog* https://www.uber.com/us/en/blog/h3/ (2018).
6. Sahr, K., White, D. & Kimerling, A. J. Geodesic Discrete Global Grid Systems. *Cartogr. Geogr. Inf. Sci.* **30**, 121–134 (2003).
7. Snyder, J. P. An Equal-Area Map Projection For Polyhedral Globes. *Cartographica* **29**, 10–21 (1992).
8. Sahr, K. https://github.com/sahrk/DGGRID). (2026).
9. Juhász, L. https://github.comgatorlab-geo/dggs-bench. GATOR Lab @ UF (2026).

10. Juhász, L. How Do Discrete Global Grid Systems Actually Perform? A Systematic Benchmark Across Geometry, Computation and Relational Joins. Preprint at https://doi.org/10.31223/X5B47J (2026).
11. Melchiorri, M *et al. Stats in the City: The GHSL Urban Centre Database 2025 : Public Release GHS UCDB R2024*. https://data.europa.eu/doi/10.2760/3046391 (2024) doi:10.2760/3046391.
12. Juhász, L. Accelerating Point-in-Polygon Predicates via Algebraic Hash-Joins and Discrete Global Grids at Scale: An Interactive Benchmark. in *The 34th ACM International Conference on Advances in Geographic Information Systems* (ACM, Riverside, CA, USA, 2026). doi:10.1145/3841645.3843101.
13. Overture. Overture Maps Foundation schema repository. https://github.com/OvertureMaps/schema (2026).