SOAR: A Scalable, Open, Automated, and Reproducible Urban Data Model for the EU

Gareth Simons[a,*], Second Author[a], Third Author[a]

[a] *Benchmark Urbanism, [TODO: Full postal address], Country*

Abstract

Spatial analytic workflows for evidence-based urban planning are typically time-intensive and data-source specific, limiting cross-context application. We present SOAR (Scalable, Open, Automated, and Reproducible), a pan-European urban data model providing standardised morphology and accessibility metrics for 699 urban centres. SOAR combines EU-specific datasets—Eurostat High Density Clusters for boundary definition, Census 2021 for demographics, and Copernicus Urban Atlas for land cover—with Overture Maps for street networks, points of interest, and building footprints. The dataset provides over 100 metrics per street network node at six spatial scales (400–9,600m), encompassing network centrality, land-use accessibility and diversity, building morphology, green space proximity, and demographics. Network processing employs automated cleaning via the cityseer Python package with level-aware processing at bridges and classification-based edge merging. Data are distributed as GeoPackage files in EPSG:3035 projection with open-source processing code. SOAR bridges the gap between global datasets with variable quality and national datasets with restrictive licensing, enabling comparative research on walkability, mixed-use development, and urban form.

Keywords: urban morphology, accessibility, European cities, walkability, open data, reproducibility, street networks

1. Introduction

Evidence-based urban design and planning (EBDP) relies on spatial analytic workflows: network analysis reveals street prominence, proximity metrics quantify access to amenities, and census data characterises population distributions. However, constructing such models is resource-intensive and data-source specific, limiting cross-context applicability. Growing interest in scalable automated models has produced datasets spanning 27,000 US street networks [1], 931

[*] Corresponding author

Email address: g.simons@benchmarkurbanism.com (Gareth Simons)

UK towns with land-use and population data [2], and 50 cities across 29 countries with network-morphology integration [3]. OpenStreetMap-derived models can scale globally but suffer variable data quality, particularly for points of interest, while global census sources lack detail. National datasets offer consistency but restrict geographic scope and often impose licensing constraints [2].

The TWIN2EXPAND consortium addresses this gap within the EU context. Continental scope necessitates broad-coverage sources (Overture Maps), while EU boundaries enable use of harmonised regional datasets:

- Eurostat 2021 Urban Centres / High Density Clusters data to rigorously define urban extents, from which we have extracted 699 towns and cities.
- Eurostat 2021 homogenised 1×1 km census statistics for population densities, employment levels, place of birth, and population change.
- Copernicus Urban Atlas data from which we derive urban blocks, building heights, and tree canopy coverage.

We use Overture Maps rather than OpenStreetMap for street networks, infrastructure, points of interest, and building footprints. Overture offers formalised release cycles, improved POI validation, and expanded building coverage via supplementary Google and Microsoft data [4]. Network processing employs automated cleaning via cityseer [5], including level-aware processing at bridges and classification-based edge merging.

SOAR (Scalable, Open, Automated, and Reproducible) demonstrates scalable methods for pan-EU urban comparison. The project prioritises openness and reproducibility, providing researchers and planners with consistent metrics for evidence-based planning and policy.

Specifications Table

Value of the Data

- These data enable **comparative urban research** across 699 European cities using consistent, standardised metrics computed from harmonised input sources, eliminating methodological inconsistencies that typically arise when combining city-specific datasets.
- The **multi-scale metric design** (400m, 800m, 1,200m, 1,600m, 4,800m, 9,600m) supports analysis at pedestrian, neighbourhood, and district scales. Based on average walking speed of 80m/min (≈5 km/h), these correspond to 5, 10, 15, 20, 60, and 120-minute walking catchments respectively—accommodating research questions from immediate walkability (5-minute neighbourhood) to metropolitan accessibility.
- Urban planners and policymakers can use these data for **benchmarking and evidence-based planning**, identifying cities with exemplary mixed-use development, green space access, or morphological characteristics to inform local planning strategies.

Subject	Geography, Urban Studies, Geospatial Data Science
Specific subject area	Urban morphology, pedestrian accessibility, land-use diversity
Type of data	Processed geospatial vector data (GeoPackage)
Data collection	Derived from publicly available authoritative sources: Overture Maps Foundation (2024), Copernicus Urban Atlas (2018), Eurostat Census Grid (2021), Copernicus Height Model (2012)
Data source location	Pan-European: 699 urban centres across EU member states. Coordinate Reference System: EPSG:3035 (ETRS89-LAEA Europe)
Data accessibility	Repository: Zenodo. Data identification number: [TODO: DOI]. URL: [TODO: URL]
Related research article	[TODO: None / or cite future CEUS paper]

- The fine-grained, node-level structure (80m street segments) provides **training data for machine learning** applications in urban pattern recognition, land-use prediction, and automated urban classification.
- Full documentation of the processing pipeline and open-source code ensures **reproducibility and extensibility**, allowing researchers to update the dataset with newer source data or adapt the methodology for other geographic regions.

2. Data Description

2.1. Dataset Structure

The dataset comprises 699 GeoPackage files, one per urban centre, following the naming convention `metrics_{bounds_fid}.gpkg` where `bounds_fid` corresponds to the unique identifier from the source high-density cluster boundaries. Each GeoPackage contains three layers:

1. **streets**: Street network nodes (80m segments) with computed metrics
2. **buildings**: Individual building footprints with morphological attributes
3. **blocks**: Urban blocks with aggregated statistics

[TODO: Insert Figure 1: Map of European coverage showing all 699 urban centres]

2.2. Streets Layer Schema

Table 1 summarises the key attributes in the streets layer. Metrics are computed at multiple distance thresholds, indicated by the suffix (e.g., `_400`, `_800`). The complete schema with all attributes is provided in Supplementary Material (Section S1).

Table 1: Summary of streets layer attributes (selected metrics shown; full schema in Supplementary Material Section S1).

Attribute group	**Description**
`cc_beta_*`	Closeness centrality (gravity-weighted) at distance thresholds
`cc_density_*`	Node density within distance threshold
`ac_{landuse}_*`	Accessibility count for land-use category
`ac_mixed_*`	Land-use diversity (Hill numbers)
`mu_hill_*`	Mixed-use indices
`green_dist_*`	Distance to nearest green space
`tree_dist_*`	Distance to nearest tree canopy
`pop_*`	Interpolated population metrics
`emp_*`	Interpolated employment metrics

2.3. Land-Use Categories

Points of interest (POIs) from Overture Maps were classified into 11 aggregated categories for accessibility analysis (Table 2). The complete mapping from Overture's 2,000+ categories to analytical classes is detailed in Supplementary Material (Section S2).

Table 2: Aggregated land-use categories derived from Overture Maps POI classification.

Category	**Includes**
`eat_and_drink`	Restaurants, cafés, bars
`retail`	Shops, markets, stores
`business_services`	Offices, professional services
`public_services`	Government, civic facilities
`health_medical`	Clinics, pharmacies, hospitals
`education`	Schools, universities, libraries
`accommodation`	Hotels, hostels
`active_life`	Sports, fitness, recreation
`arts_entertainment`	Cinemas, theatres, venues
`attractions`	Museums, landmarks, parks
`religious`	Places of worship

2.4. Buildings and Blocks Layers

The buildings layer contains individual building footprints with morphological metrics including area, perimeter, compactness, orientation, estimated height (sampled from raster), and floor area ratio. The blocks layer provides urban block geometries derived from Urban Atlas with aggregated building coverage ratios and block-level morphology.

[TODO: Insert Figure 2: Sample visualisation of metrics for one city (e.g., Barcelona or Vienna)]

3. Experimental Design, Materials and Methods

3.1. Study Area Definition

Urban boundaries derive from the Eurostat High Density Clusters (HDENS-CLST) 2021 raster [6]—a 1×1 km grid identifying contiguous cells with ≥1,500 inhabitants/km^2 and cumulative population ≥50,000. Vectorisation, filtering to continental Europe (EPSG:3035), and UK exclusion yielded 699 urban centres. City names were assigned via spatial join with Overture Maps administrative divisions.

3.2. Data Sources

Table 3 summarises input datasets. Street networks from Overture Maps [4] were cleaned via cityseer [5]: disconnected components removed, degree-2 nodes consolidated, edges decomposed to 80m segments (≈1-minute walk). Level-aware processing prevents incorrect merging at bridges. POI categories from Overture's "places" theme were mapped to 11 analytical classes (Table 2). Land cover and tree canopy derive from Copernicus Urban Atlas [7] and Street Tree Layer [8] respectively. Building heights were sampled from the Digital Height Model [9]. Demographics from Eurostat Census Grid 2021 [10] include population, employment, age structure, nationality, and migration at 1 km^2 resolution.

Table 3: Input data sources and their roles in the processing pipeline.

Dataset	**Source**	**Use**
Street networks	Overture Maps	Centrality, accessibility
Points of interest	Overture Maps	Land-use metrics
Building footprints	Overture Maps	Morphology
Urban Atlas 2018	Copernicus	Blocks, green space
Street Tree Layer	Copernicus	Tree proximity
Height Model 2012	Copernicus	Building heights
Census Grid 2021	Eurostat	Demographics

3.3. Data Processing Pipeline

The pipeline comprises six stages: (1) boundary vectorisation from HDENS-CLST raster; (2–4) clipping of Copernicus datasets (Urban Atlas, Street Tree Layer, Height Model) to boundaries; (5) Overture data extraction per boundary; and (6) metric computation. All scripts support idempotent execution—existing outputs are skipped unless `-overwrite` is specified—enabling incremental processing and failure recovery.

To ensure accurate metric computation at boundary edges, input datasets are buffered beyond city boundaries by their respective maximum aggregation distances. Street networks are buffered by 9,600m (the maximum centrality threshold), ensuring that nodes near boundaries have complete network context for centrality calculations. Points of interest are buffered by 1,600m (the maximum pedestrian-scale accessibility threshold), capturing all reachable amenities for boundary-proximate locations. Copernicus land cover and tree canopy data are similarly buffered to ensure complete coverage for green space proximity metrics.

3.4. Metric Computation

3.4.1. Network Centrality

Network centrality metrics quantify a location's prominence within the street network and were computed using the cityseer package [5]. The segment-based approach computes centrality relative to reachable street segments within distance thresholds, avoiding distortions from irregular node distributions. Beta-weighted (gravity index) closeness applies an explicit decay parameter β:

$$C_i^{\beta} = \sum_{j \neq i} w_j \cdot e^{-\beta \cdot d_{ij}} \tag{1}$$

where d_{ij} is network distance, w_j is segment length, and β controls distance decay [5]. Harmonic closeness—appropriate for localised implementations constrained by threshold $d_{\max}$—was also computed. Metrics were generated at thresholds of 400m, 800m, 1,200m, 1,600m, 4,800m, and 9,600m, corresponding to 5, 10, 15, 20, 60, and 120-minute walking catchments at 80m/min average pedestrian speed. The shorter thresholds (5–20 minutes) capture pedestrian-scale accessibility relevant to daily errands and commuting choices, while longer thresholds (60–120 minutes) characterise district and metropolitan-scale network structure.

3.4.2. Land-Use Accessibility

Accessibility metrics aggregate POI counts over the street network from identical locations as centrality computations, enabling direct correlation [5]. For each land-use category k, both unweighted counts (number of reachable POIs within $d_{\max}$) and distance-weighted counts (applying exponential decay) were computed, alongside nearest-distance measures.

Mixed-use diversity was quantified using Hill numbers [5], the preferred diversity index because it adheres to the replication principle and uses units of effective species:

$$ {}^{q}D = \left(\sum_{k=1}^{K} p_k^q \right)^{1/(1-q)} \tag{2} $$

where p_k is the proportion of POIs in category k and q controls sensitivity to rare categories. At $q = 0$, Hill numbers reduce to a simple count of distinct land-use types (species richness); at $q = 1$, they approximate the exponential of Shannon entropy; at $q = 2$, they approximate the inverse Simpson index, emphasising balance over richness. Following cityseer recommendations, we compute $q = 0$, $q = 1$, and $q = 2$ variants in both unweighted and distance-weighted forms.

3.4.3. Morphology

Building morphology metrics were computed using standard geometric formulae: compactness as $4\pi A/P^2$, orientation via minimum bounding rectangle, and form factor as $A/(h \cdot P)$ where h is height.

3.4.4. Green Space Proximity

Distance to nearest green space polygon edge was computed for each network node using spatial indexing.

3.4.5. Demographic Interpolation

Census grid values were interpolated to network nodes using linear interpolation from grid cell centroids.

3.5. Implementation

The pipeline is implemented in Python 3.12 with four modules: data ingestion (`src.data`), metric processing (`src.processing`), utilities (`src.tools`), and land-use classification (`src.landuse_categories`).

Data loading abstracts heterogeneous sources through a unified interface. Network loading retrieves Overture “connector” and “segment” themes via STAC, transforms to EPSG:3035, constructs a NetworkX MultiGraph with node deduplication and edge attribute preservation, then applies cityseer cleaning (8m tolerance, 100-node component threshold). Building and infrastructure loading extract footprints and point features respectively. POI loading maps Overture’s 2,000+ categories to 23 intermediate classes via CSV lookup, then to 11 analytical categories.

Land-use classification consolidates Overture’s taxonomy in two stages: schema filtering retains 23 intermediate classes, then category merging yields 11 analytical categories (Table 2) aligned with established urban taxonomies.

The processing module (`generate_metrics.py`) orchestrates metric computation across all boundaries. Networks undergo dual graph transformation (streets as nodes, intersections as edges) after 80m decomposition, enabling segment-level analysis.

Centrality: Beta-weighted (gravity index) and harmonic closeness at six thresholds (400m, 800m, 1,200m, 1,600m, 4,800m, 9,600m) via cityseer [5].

Accessibility: POI counts (unweighted and distance-weighted) per land-use category within pedestrian-scale thresholds (400m–1,600m, i.e., 5–20-minute walks); Hill number diversity indices ($q = 0, 1, 2$) quantifying land-use heterogeneity [5].

Morphology: Building metrics (area, compactness, orientation, height, volume) and block metrics (coverage ratio, shape) computed via momepy, then aggregated to network nodes.

Green space: Network-distance proximity to Urban Atlas green classes and Street Tree Layer polygons, using point-sampled boundaries at 20m intervals.

Demographics: Census grid values interpolated to nodes via linear interpolation from cell centroids.

3.6. Output Format

Results are exported as GeoPackage files (`metrics_{bounds_fid}.gpkg`) in EPSG:3035, each containing three layers: `streets` (network nodes with all computed metrics), `buildings` (footprints with morphology), and `blocks` (polygons with coverage ratios). Dependencies are pinned via `pyproject.toml`; key packages include cityseer (network analysis), momepy (morphometrics), geopandas (spatial data), and overturemaps (data access). Processing parameters are fixed for reproducibility: 80m decomposition (≈1-minute walk), 8m cleaning tolerance, centrality thresholds [400, 800, 1200, 1600, 4800, 9600]m (5–120 minutes), accessibility thresholds [400, 800, 1200, 1600]m (5–20 minutes).

3.7. Potential Applications

SOAR's standardized multi-scale metrics enable diverse research applications across urban planning, geography, and data science. We present ten illustrative examples demonstrating the breadth of potential issues and questions that can be explored in derivative research based on this dataset:

1. **POI Data Quality Assessment**: Use Random Forest regression to predict expected POI counts from population; compute z-scores to identify cities with systematic data gaps; aid filtering of dataset to reliable cities before conducting downstream analyses.
2. **Green Space Accessibility Equity**: Measure distance-weighted access to green blocks and tree canopies; stratify by census demographics within cities to identify environmental justice gaps.
3. **Educational Infrastructure Gap Analysis**: Overlay walking catchments to educational facilities; overlay with child population from census grids to flag underserved areas.
4. **ML-Based POI Demand Modeling**: Train models using population density and network centrality to predict restaurant/cafe density; identify undersupplied neighborhoods.
5. **15-Minute City Benchmarking**: Identify streets with access to all 11 POI categories within 15-minute (1200m) and 20-minute (1600m) walks; rank cities by completeness.

6. **Comparative Urban Density Patterns**: Compare population density and building density morphology distributions across cities.
7. **Pan-European Amenity Distance Atlas**: Compute median walking distances to essential services per city; aggregate by country (where sufficient coverage) to reveal North-South and East-West gradients.
8. **Mixed-Use Development Scoring**: Combine Hill diversity indices with building morphology metrics to classify neighborhoods along single-use ↔ mixed-use spectrum.
9. **Population Intensification Opportunities**: Identify high-centrality, high-diversity nodes with low current population density as optimal densification targets.
10. **Transit Station Gap Analysis**: Map density against existing station locations to identify underserved high-demand areas.

These examples illustrate SOAR's versatility for equity analysis, infrastructure planning, benchmarking, predictive modeling, and comparative geography. Prototype methodological workflows for each application are presented in a companion research article [TODO: cite demonstrators paper].

Ethics Statement

This research did not involve human subjects, animal experiments, or data collected from social media platforms. All source datasets are publicly available under open licenses.

CRediT Author Statement

Gareth Simons: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Visualization. [TODO: Add other authors and their contributions]

Declaration of Competing Interest

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

Acknowledgements

[TODO: Acknowledge TWIN2EXPAND project, funding bodies, and any individuals who assisted.]

This work was supported by [TODO: the European Union's Horizon programme under grant agreement No. XXXXX (TWIN2EXPAND project)].

Data Availability and Code Availability

The dataset is available at Zenodo: [TODO: Insert DOI and URL after deposit].

Supplementary Material accompanies this article and includes:

- Section S1: Complete streets layer schema with all ∼100 attributes
- Section S2: Full land-use classification mapping from Overture categories to analytical classes
- Section S3: Detailed metadata for all source datasets (HDENS-CLST, Overture Maps, Copernicus Urban Atlas, Street Tree Layer, Building Height, Census Grid)
- Section S4: Processing parameters (network cleaning, centrality computation, accessibility thresholds, POI saturation assessment)
- Section S5: Software dependencies with pinned versions
- Section S6: Computational requirements and parallelisation strategies
- Section S7: Data quality notes, known limitations, and QA procedures
- Section S8: Citation information for dataset, software, and source data

The processing workflow is implemented in the *ebdptoolkit* repository (version 0.5.0), available at `https://github.com/UCL/ba-ebdp-toolkit`. The toolkit provides a complete, reproducible pipeline for generating urban metrics from open data sources (Overture Maps, Copernicus Urban Atlas, and census grids). The workflow comprises four stages: boundary generation from raster clusters, data ingestion from distributed sources, metric computation via network and morphological analysis, and POI saturation assessment via multi-scale regression.

To use the workflow: (1) clone the repository and install dependencies using `uv sync`; (2) download input datasets (boundaries raster, Overture dumps, Urban Atlas shapefiles) as documented in the README; (3) run `python -m src.processing.generate_metrics` with parameters specifying output directory and boundary file; (4) optionally run `src/analysis/poi_saturation_notebook.py` to assess data completeness and generate quadrant classifications. All processing parameters (network decomposition, cleaning thresholds, distance scales) are configurable via function arguments. Complete documentation, data loading guidelines, and usage examples are provided in the repository's README and inline code comments. The toolkit is licensed under AGPL-3.0 and depends on open-source packages (cityseer, momepy, geopandas) ensuring full transparency and reproducibility across environments.

References

1. G. Boeing, A multi-scale analysis of 27,000 urban street networks: Every US city, town, urbanized area, and Zillow neighborhood, Environ. Plan. B Urban Anal. City Sci. 47 (4) (2020) 590–608. https://doi.org/10.1177/2399808318784595

2. G.D. Simons, Detection and prediction of urban archetypes at the pedestrian scale: computational toolsets, morphological metrics, and machine learning methods, Ph.D. thesis, UCL (University College London), 2021. https://discovery.ucl.ac.uk/id/eprint/10134012/

3. W. Yap, F. Biljecki, A global feature-rich network dataset of cities and dashboard for comprehensive urban analyses, Sci. Data 10 (2023) 667. https://doi.org/10.1038/s41597-023-02578-1

4. Overture Maps Foundation, Overture Maps Data – 2024 Release [dataset], 2024. https://overturemaps.org/

5. G. Simons, The cityseer Python package for pedestrian-scale network-based urban analysis, Environ. Plan. B Urban Anal. City Sci. 49 (9) (2022) 2356–2361. https://doi.org/10.1177/23998083221133827

6. European Commission, Joint Research Centre, High Density Clusters – HDENS-CLST 2021 [dataset], 2023. https://ghsl.jrc.ec.europa.eu/

7. European Environment Agency, Urban Atlas 2018 [dataset], Copernicus Land Monitoring Service, 2020. https://land.copernicus.eu/local/urban-atlas

8. European Environment Agency, Street Tree Layer 2018 [dataset], Copernicus Land Monitoring Service, 2020. https://land.copernicus.eu/local/urban-atlas/street-tree-layer-stl-2018

9. European Environment Agency, Building Height 2012 [dataset], Copernicus Land Monitoring Service, 2020. https://land.copernicus.eu/local/urban-atlas/building-height-2012

10. Eurostat, Census 2021 – Population Grid [dataset], 2024. https://ec.europa.eu/eurostat/web/gisco/geodata/population-distribution/geostat