

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

Constructing spatial analytic datasets for urban planning is resource-intensive and data-source specific, limiting cross-context application. We present SOAR (Scalable, Open, Automated, and Reproducible), a pan-European urban data model providing 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 multiple spatial scales (100–4,800m), encompassing network centrality, land-use accessibility and diversity, building morphology, green space proximity, and demographics. Processing is automated via open-source Python code with fixed parameters for reproducibility. Data are distributed as GeoPackage files in EPSG:3035 projection. This is a derived dataset from open sources processed at continental scale; it is not a curated or validated dataset, and users should assess data quality for their specific applications—particularly for point of interest data, which exhibits geographic variation in completeness. The primary contribution is reducing the barrier to pan-European urban comparison by providing pre-computed metrics in a consistent format.

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

1. Introduction

Constructing spatial analytic datasets for urban planning requires substantial data engineering: network analysis, proximity metrics, census integration, and land-use classification. This work is typically repeated for each study area. Existing large-scale datasets include 27,000 US street networks [1], 931 UK towns [2], and 50 cities across 29 countries [3], but pan-European coverage with consistent methodology has been limited.

The TWIN2EXPAND consortium produced SOAR to address this gap. Continental scope requires 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.

*Corresponding author

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

SOAR (Scalable, Open, Automated, and Reproducible) provides pre-computed urban metrics for pan-EU comparison. The processing pipeline is open-source and reproducible. The primary contribution is the operational scale: deriving consistent metrics across 699 urban centres from heterogeneous open data sources simplifies downstream comparative work that would otherwise require substantial data engineering effort.

Specifications Table

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]

Value of the Data

- **Consistent metrics across 699 cities:** Metrics are computed using identical methods and parameters, avoiding methodological inconsistencies that arise when combining city-specific datasets.
- **Multi-scale design:** Centrality metrics at 400m, 800m, 1,200m, 1,600m, and 4,800m (5–60 minute walking catchments). Accessibility metrics at 200m, 400m, 800m, 1,200m, and 1,600m.
- **Network-based aggregation:** Metrics computed along pedestrian routes rather than Euclidean buffers. A park 200m as-the-crow-flies but 800m by foot (due to barriers) is measured at 800m.
- **Distance-weighted accessibility variants:** Accessibility metrics are provided in both unweighted and distance-weighted forms. The distance-weighted metrics use an Exponential decay ($e^{-\beta d}$) function that weights nearby POIs more heavily than distant ones.
- **Fine-grained resolution:** 80m street segments preserve within-neighbourhood heterogeneity.
- **Open and reproducible:** Processing code and parameters are documented; researchers can regenerate or extend the dataset.

Example Use Cases

Table 1 illustrates potential applications of SOAR’s pre-computed metrics. A companion paper [\[TODO: cite demonstrators paper\]](#) provides worked examples for each use case.

Table 1: Example research applications enabled by SOAR.

Use Case	Relevant Metrics	Example Application
Data quality filtering	POI counts vs. population density	Identify cities with sparse POI coverage via regression residuals
Multi-scale analysis	Metrics at multiple thresholds	Compare within-city vs. between-city relationships
Access gap identification	Distance-to-nearest metrics	Locate areas with poor access to education or transport
Predictive modelling	Centrality, density, accessibility	Train cross-city models for land-use intensity prediction
Benchmarking	All accessibility metrics	Rank cities by walkable access to amenities
Typology classification	Morphology, centrality, mixed-use	Cluster neighbourhoods into recurring urban forms
Site selection	Centrality, accessibility, density	Filter locations meeting multiple development criteria

2. Data Description

2.1. Dataset Structure

The dataset comprises a `boundaries.gpkg` file extracted from the source high-density cluster boundaries and a ZIP file containing 699 GeoPackage files, one per urban centre defined by the boundaries, following the naming convention `metrics_{bounds_fid}.gpkg` where `bounds_fid` corresponds to the unique identifier from `boundaries.gpkg`. Each derivative 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 source attributes plus morphological attributes

2.2. Streets Layer Schema

Table 2 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).

2.3. Land-Use Categories

Points of interest (POIs) from Overture Maps were classified into 11 aggregated categories for accessibility analysis (Table 3). The complete mapping from Overture’s 2,000+ categories to analytical classes is detailed in Supplementary Material (Section S2).

Infrastructure points from Overture Maps were separately classified into three analytical categories for transport and amenity accessibility (Table 4).

2.4. Buildings and Blocks Layers

The buildings layer contains individual building footprints derived from Overture Maps 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.

Table 2: 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_{landuse}_*_nw	Unweighted POI count for land-use category
cc_{landuse}_*_wt	Distance-weighted POI count for land-use category
cc_{landuse}_nearest_*	Distance to nearest POI in category
cc_hill_*	Land-use diversity (Hill numbers $q = 0, 1, 2$)
cc_green_nearest_*	Distance to nearest green space
cc_trees_nearest_*	Distance to nearest tree canopy
cc_*_median_*	Morphology metrics (median aggregation)
density, t, emp	Interpolated census metrics

Table 3: Aggregated land-use categories derived from Overture Maps POI classification.

Category	Includes
eat_and_drink	Restaurants, cafés, bars
retail	Shops, markets, stores
business_and_services	Automotive, beauty/spa, financial, professional services, real estate, travel agencies
public_services	Government offices, civic facilities
health_and_medical	Clinics, pharmacies, hospitals
education	Schools, universities, libraries
accommodation	Hotels, hostels, lodging
active_life	Sports facilities, fitness centres, recreation
arts_and_entertainment	Cinemas, theatres, performance venues
attractions_and_activities	Museums, landmarks, tourist attractions
religious	Places of worship

Table 4: Infrastructure categories derived from Overture Maps infrastructure theme.

Category	Includes
transport	Bus stops, bus stations, railway stations, subway stations, ferry terminals, airports
street_furn parking	Benches, drinking fountains, planters, post boxes, picnic tables Car parking, motorcycle parking

3. 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 $\geq 1,500$ inhabitants/km² and cumulative population $\geq 50,000$. 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 5 summarises input datasets. Street networks from Overture Maps [4] were cleaned via cityseer [5]: disconnected components removed, degree-2 nodes removed, edges decomposed to 80m segments (≈ 1 -minute walk). Level-aware processing prevents incorrect merging at bridges or tunnels. POI categories from Overture’s “places” theme were mapped to 11 analytical classes (Table 3). 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 km² resolution.

Table 5: 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 executed via Python modules (Table 6). All scripts support idempotent execution—existing outputs are skipped unless `overwrite` is specified—enabling incremental processing and failure recovery.

Table 6: Processing pipeline stages with corresponding Python modules.

Stage	Module	Description
1	<code>src.data.generate_boundary_polys</code>	Vectorise HDENS-CLST raster to boundary polygons
2	<code>src.data.load_urban_atlas_blocks</code>	Extract Urban Atlas blocks per boundary
3	<code>src.data.load_urban_atlas_trees</code>	Extract Street Tree Layer per boundary
4	<code>src.data.load_bldg_hts_raster</code>	Clip building height rasters per boundary
5	<code>src.data.load_overture</code>	Extract Overture data (network, POIs, buildings, infrastructure)
6	<code>src.processing.generate_metrics</code>	Compute all metrics per boundary

To ensure accurate metric computation at boundary edges, input datasets are buffered beyond city boundaries. Street networks are buffered by 10km, ensuring that nodes near boundaries have complete network context for centrality calculations. Points of interest, buildings, and infrastructure are buffered by 2km, capturing all reachable amenities for boundary-proximate locations. Copernicus land cover and tree canopy data are clipped to boundary bounding boxes to ensure 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. Closeness variants include: beta-weighted (gravity index with exponential distance decay), harmonic (sum of inverse distances), Hillier-type (node count divided by average node distance), farness (sum of distances), and simple cycles and street density counts. Betweenness centrality—measuring how often a node lies on shortest paths between other nodes—is computed in both standard and distance-weighted forms. Metrics were generated at thresholds of 400m, 800m, 1,200m, 1,600m, and 4,800m, corresponding to 5, 10, 15, 20, and 60-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 the 60-minute threshold characterises district-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]. Unlike Euclidean buffer approaches or zonal aggregation (which assign uniform values to all locations within administrative units), network-based accessibility follows actual pedestrian routes and varies continuously across street segments. 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. Distance-weighted counts better reflect perceived accessibility but at the cost of the simpler interpretability of unweighted counts: a café at 100m contributes more than one at 1,000m, reflecting that amenity usage decays with distance.

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:

$${}^qD = \left(\sum_{k=1}^K p_k^q \right)^{1/(1-q)} \quad (1)$$

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 and block morphology metrics were computed using the momepy package [11], which provides standardised urban morphometric functions. For each building footprint, we compute: area, perimeter, circular compactness, orientation, corners count, shape index, fractal dimension, and shared wall length with adjacent buildings. Building heights are sampled from the Copernicus Height Model raster, enabling computation of volume, floor area ratio (assuming 3m floor height), and form factor. For urban blocks derived from Urban Atlas, we compute: area, perimeter, circular compactness, orientation, and building coverage ratio (sum of building areas within block divided by block area).

Building and block metrics are aggregated to network nodes via nearest-network-distance assignment within 100m and 200m thresholds. Each building or block centroid is assigned to the nearest network node, and statistics (median and median absolute deviation) are computed across all features within each threshold distance in both unweighted and distance-weighted forms. This approach captures local morphological context while preserving spatial heterogeneity across the street network.

3.4.4. Green Space Proximity

Distance to nearest green space polygon edge was computed for each network node. Green spaces are identified from Urban Atlas land cover classes: green urban areas, forests, pastures, arable land, orchards, permanent crops, sports and leisure facilities, herbaceous vegetation, wetlands, water bodies, and open spaces

with little vegetation. Polygon boundaries are sampled at 20m intervals to generate point features for network-distance calculations to improve proximity accuracy for large or complex polygons, since large polygons can be approached from many directions (only the first instance of each original polygon’s boundary points is considered per catchment). Tree canopy proximity uses Street Tree Layer polygons with equivalent sampling. Both green space and tree canopy areas are aggregated within walking distances (200m, 400m, 800m) to quantify local green coverage.

3.4.5. Demographic Interpolation

Census grid values were interpolated to network nodes using linear interpolation from grid cell centroids. Interpolated variables include: total population (**t**), population density (**density**), male/female counts (**m**, **f**), age cohorts (under 15, 15–64, 65+), employment (**emp**), nationality (national, EU other, non-EU), and migration flows (same residence, in-migration, out-migration). Percentage variants are computed for each variable relative to total population.

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**). Network loading retrieves Overture “connector” and “segment” themes via STAC, transforms to EPSG:3035, constructs a NetworkX MultiGraph, then applies cityseer cleaning (8m tolerance, 100-node minimum component size). Networks undergo dual graph transformation (streets as nodes, intersections as edges) after 80m decomposition (\approx 1-minute walk), enabling segment-level analysis. POI loading maps Overture’s 2,000+ categories to 23 intermediate classes via CSV lookup, then to 11 analytical categories (Table 3).

Table 7 summarises the fixed processing parameters ensuring reproducibility across all urban centres.

Table 7: Fixed processing parameters for reproducibility.

Parameter	Value(s)	Rationale
Network decomposition	80m	\approx 1-minute walking segment
Network cleaning tolerance	8m	Consolidate near-parallel edges
Min. component size	100 nodes	Remove disconnected fragments
Centrality thresholds	400, 800, 1200, 1600, 4800m	5–60 minute walking catchments
Accessibility thresholds	200, 400, 800, 1200, 1600m	Pedestrian-scale proximity
Morphology aggregation	100, 200m	Immediate building context
Green space proximity	1600m	Extended walkable distance
Green area aggregation	200, 400, 800m	Multi-scale green coverage

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), momapy (morphometrics), geopandas (spatial data), and overturemaps (data access). All processing parameters are fixed for reproducibility (Table 7).

3.7. Data Quality Considerations

SOAR is a derived dataset processed via automated pipeline from open data sources. It is not a curated or validated dataset. Users should assess fitness for purpose before analysis:

- **Point of interest data:** POI completeness varies geographically. Some Central and Western European cities (Germany, Netherlands) may exhibit high saturation coverage, while peripheral regions (parts of Spain, Romania, Bulgaria, southern Italy) may show systematic undersaturation. A companion paper [TODO: cite demonstrators] provides a multi-scale regression approach for assessing POI saturation per city.

- **Building heights:** Derived from 2012 Copernicus raster data; heights may be incomplete or may be outdated in areas with recent construction.
- **Street networks:** Automated cleaning removes disconnected components and simplifies topology; some local idiosyncrasies or redundancies may persist.
- **Census interpolation:** Demographics are interpolated from 1 km² grid cells to network nodes; this smooths within-cell variation.

The dataset is suitable for comparative analysis across cities at aggregate scales. Fine-grained analysis (e.g., individual street segments in undersaturated cities) may require additional validation against local data sources.

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.

AI

The original code was developed directly by the authors with minimal AI assistance. Later reviews, edits, and preparation of the vignettes used AI to more rapidly generate and refine the content. This manuscript was prepared with the assistance of AI tools for content, language, editing, and formatting. The authors reviewed and verified all content to ensure accuracy and integrity.

CRedit Author Statement

Gareth Simons: Conceptualisation, Methodology, Software, Data curation, Writing – original draft, Visualization.

Sepehr Zhand & Kayvan Karimi: Conceptualisation, review, and editing.

Acknowledgements

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 *t2e-soar* repository, available at <https://github.com/benchmark-urbanism/t2e-soar>. 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 six stages (Table 6): boundary generation from raster clusters, Urban Atlas and tree canopy extraction, building height raster clipping, Overture data ingestion, and metric computation.

To use the workflow: (1) clone the repository and install dependencies using `uv sync`; (2) download input datasets (HDENS-CLST raster, Urban Atlas shapefiles, building height rasters, census grid) as documented in the README; (3) run the data loading scripts in sequence (Table 6); (4) run `python -m src.processing.generate_metrics` with parameters specifying input paths and output directory. Preparation of the datasets, including Overture Maps download and network cleaning, takes approximately 1 to 2 days and generation of derivative metrics takes approximately 3-4 days (M1 Mac). 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>
11. M. Fleischmann, momepy: Urban morphology measuring toolkit, *J. Open Source Softw.* 4 (43) (2019) 1807. <https://doi.org/10.21105/joss.01807>