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

Contents

1 Complete Streets Layer Schema

The streets layer contains network nodes with computed metrics. All distance-dependent metrics are computed at multiple thresholds: 400m, 800m, 1200m, 1600m (for accessibility), and additionally 4800m, 9600m (for centrality). Table 1 provides the complete attribute schema.

Table 1: Complete streets layer attribute schema. Suffix _XXX indicates distance threshold in metres.

Attribute	Type	Description	Source/Method
Geometry and Identifiers			
geometry	Point	Node location (EPSG:3035)	Network decomposition
node_id	Integer	Unique node identifier	Sequential assignment
x	Float	X coordinate (EPSG:3035)	Geometry
y	Float	Y coordinate (EPSG:3035)	Geometry
Network Centrality (Segment-Based)			
cc_beta_XXX	Float	Beta-weighted closeness (gravity index)	cityseer segment centrality
cc_harmonic_XXX	Float	Harmonic closeness	cityseer segment centrality
cc_density_XXX	Float	Node density within threshold	cityseer segment centrality
cc_farness_XXX	Float	Total network distance to reachable nodes	cityseer segment centrality
cc_betweenness_XXX	Float	Segment betweenness centrality	cityseer segment centrality
Accessibility – POI Counts (per land-use category)			
ac_accommodation_XXX_nw	Integer	Unweighted count of accommodation POIs	cityseer accessibility
ac_accommodation_XXX_wt	Float	Distance-weighted count of accommodation POIs	cityseer accessibility
ac_accommodation_nearest_max_XXX	Float	Distance to nearest accommodation POI (m)	cityseer accessibility
ac_active_life_XXX_nw	Integer	Unweighted count of active life POIs	cityseer accessibility
ac_active_life_XXX_wt	Float	Distance-weighted count of active life POIs	cityseer accessibility
ac_active_life_nearest_max_XXX	Float	Distance to nearest active life POI (m)	cityseer accessibility

Continued on next page

Table 1 – continued from previous page

Attribute	Type	Description	Source/Method
ac_arts_entertainment_XXX_nw	Integer	Unweighted count of arts/entertainment POIs	cityseer accessibility
ac_arts_entertainment_XXX_wt	Float	Distance-weighted count of arts/entertainment POIs	cityseer accessibility
ac_arts_entertainment_nearest_max_XXX	Float	Distance to nearest arts/entertainment POI (m)	cityseer accessibility
ac_attractions_XXX_nw	Integer	Unweighted count of attractions POIs	cityseer accessibility
ac_attractions_XXX_wt	Float	Distance-weighted count of attractions POIs	cityseer accessibility
ac_attractions_nearest_max_XXX	Float	Distance to nearest attractions POI (m)	cityseer accessibility
ac_business_services_XXX_nw	Integer	Unweighted count of business/services POIs	cityseer accessibility
ac_business_services_XXX_wt	Float	Distance-weighted count of business/services POIs	cityseer accessibility
ac_business_services_nearest_max_XXX	Float	Distance to nearest business/services POI (m)	cityseer accessibility
ac_eat_and_drink_XXX_nw	Integer	Unweighted count of eat/drink POIs	cityseer accessibility
ac_eat_and_drink_XXX_wt	Float	Distance-weighted count of eat/drink POIs	cityseer accessibility
ac_eat_and_drink_nearest_max_XXX	Float	Distance to nearest eat/drink POI (m)	cityseer accessibility
ac_education_XXX_nw	Integer	Unweighted count of education POIs	cityseer accessibility
ac_education_XXX_wt	Float	Distance-weighted count of education POIs	cityseer accessibility
ac_education_nearest_max_XXX	Float	Distance to nearest education POI (m)	cityseer accessibility
ac_health_medical_XXX_nw	Integer	Unweighted count of health/medical POIs	cityseer accessibility
ac_health_medical_XXX_wt	Float	Distance-weighted count of health/medical POIs	cityseer accessibility

Continued on next page

Table 1 – continued from previous page

Attribute	Type	Description	Source/Method
ac_health_medical_nearest_max_XXX	Float	Distance to nearest health/medical POI (m)	cityseer accessibility
ac_public_services_XXX_nw	Integer	Unweighted count of public services POIs	cityseer accessibility
ac_public_services_XXX_wt	Float	Distance-weighted count of public services POIs	cityseer accessibility
ac_public_services_nearest_max_XXX	Float	Distance to nearest public services POI (m)	cityseer accessibility
ac_religious_XXX_nw	Integer	Unweighted count of religious POIs	cityseer accessibility
ac_religious_XXX_wt	Float	Distance-weighted count of religious POIs	cityseer accessibility
ac_religious_nearest_max_XXX	Float	Distance to nearest religious POI (m)	cityseer accessibility
ac_retail_XXX_nw	Integer	Unweighted count of retail POIs	cityseer accessibility
ac_retail_XXX_wt	Float	Distance-weighted count of retail POIs	cityseer accessibility
ac_retail_nearest_max_XXX	Float	Distance to nearest retail POI (m)	cityseer accessibility
Mixed-Use Diversity Indices			
mu_hill_q0_XXX_nw	Float	Hill diversity (q=0, richness) unweighted	cityseer mixed-use
mu_hill_q0_XXX_wt	Float	Hill diversity (q=0, richness) distance-weighted	cityseer mixed-use
mu_hill_q1_XXX_nw	Float	Hill diversity (q=1, exp Shannon) unweighted	cityseer mixed-use
mu_hill_q1_XXX_wt	Float	Hill diversity (q=1, exp Shannon) distance-weighted	cityseer mixed-use
mu_hill_q2_XXX_nw	Float	Hill diversity (q=2, inv Simpson) unweighted	cityseer mixed-use
mu_hill_q2_XXX_wt	Float	Hill diversity (q=2, inv Simpson) distance-weighted	cityseer mixed-use
Building Morphology (aggregated from buildings layer)			
bldg_area_mean_XXX	Float	Mean building area within threshold (m^2)	momepy + aggregation
bldg_area_sum_XXX	Float	Total building area within threshold (m^2)	momepy + aggregation
bldg_perimeter_mean_XXX	Float	Mean building perimeter (m)	momepy + aggregation

Continued on next page

Table 1 – continued from previous page

Attribute	Type	Description	Source/Method
bldg_compactness_mean_XXX	Float	Mean compactness ($4\pi A/P^2$)	momepy + aggregation
bldg_orientation_mean_XXX	Float	Mean orientation (degrees)	momepy + aggregation
bldg_height_mean_XXX	Float	Mean building height (m)	Copernicus DHM + aggregation
bldg_volume_mean_XXX	Float	Mean building volume (m^3)	Area × Height
bldg_floor_area_ratio_XXX	Float	Floor area ratio (total floor area / land area)	Aggregation
Green Space Proximity			
green_dist_XXX	Float	Distance to nearest green space (m)	Urban Atlas + spatial join
green_area_XXX	Float	Total green space area within threshold (m^2)	Urban Atlas + aggregation
tree_dist_XXX	Float	Distance to nearest tree canopy (m)	Street Tree Layer + spatial join
tree_area_XXX	Float	Total tree canopy area within threshold (m^2)	Street Tree Layer + aggregation
Demographics (interpolated from census grid)			
pop_total	Float	Total population (interpolated)	Census 2021 + interpolation
pop_density_XXX	Float	Population density within threshold (per km^2)	Census 2021 + aggregation
pop_change	Float	Population change 2011-2021 (%)	Census 2021
emp_total	Float	Total employment (interpolated)	Census 2021 + interpolation
emp_density_XXX	Float	Employment density within threshold (per km^2)	Census 2021 + aggregation
age_0_14_pct	Float	Population aged 0-14 (%)	Census 2021
age_15_64_pct	Float	Population aged 15-64 (%)	Census 2021
age_65_plus_pct	Float	Population aged 65+ (%)	Census 2021
foreign_born_pct	Float	Foreign-born population (%)	Census 2021
Block Characteristics (aggregated from blocks layer)			

Continued on next page

Table 1 – continued from previous page

Attribute	Type	Description	Source/Method
block_coverage_ratio_XXX	Float	Building coverage ratio within blocks	Urban Atlas + aggregation
block_area_mean_XXX	Float	Mean block area (m^2)	Urban Atlas + aggregation

2 Complete Land-Use Classification Schema

Table 2 provides the complete mapping from Overture Maps categories to the 11 analytical land-use classes.

Table 2: Complete land-use classification schema mapping Overture categories to analytical classes.

Analytical Class	Intermediate Class	Overture Categories (examples)
eat_and_drink	restaurant	restaurant, cafe, bistro, diner
	bar	bar, pub, nightclub, lounge
	coffee	coffee_shop, tea_house
retail	retail	shop, store, boutique, market
	shopping	shopping_mall, shopping_center
	grocery	supermarket, grocery_store, convenience_store
	specialty	bookshop, florist, pharmacy (retail)
business_services	office	office, coworking_space
	professional_services	law_firm, consulting, accounting, real_estate
public_services	government	town_hall, courthouse, embassy
	civic	post_office, library (public), community_center
health_medical	healthcare	hospital, clinic, medical_center
	pharmacy	pharmacy (medical), drugstore
	wellness	dentist, physiotherapy, optician
education	school	primary_school, secondary_school
	university	university, college, research_institute
	training	language_school, driving_school, training_center
accommodation	accommodation	hotel, hostel, motel, guest_house, bed_and_breakfast
active_life	sports	gym, fitness_center, sports_club, stadium
	recreation	swimming_pool, tennis_court, playground
	outdoor	park (active), hiking_trail, sports_field
arts_entertainment	culture	cinema, theatre, concert_hall, opera_house
	entertainment	bowling, arcade, casino, nightlife_venue
attractions	tourism	museum, gallery, monument, landmark
	attractions	zoo, aquarium, amusement_park, viewpoint
	nature	botanical_garden, nature_reserve, beach
religious	religious	church, mosque, synagogue, temple, monastery

3 Dataset Metadata

3.1 Eurostat High Density Clusters (HDENS-CLST) 2021

- **Source**: European Commission Joint Research Centre (JRC)
- **URL**: `https://ghsl.jrc.ec.europa.eu/`
- **Resolution**: 1km × 1km grid
- **Definition**: Contiguous cells with ≥1,500 inhabitants/km^2 and cumulative population ≥50,000
- **Reference Year**: 2021
- **Coverage**: Pan-European

- **License**: CC BY 4.0
- **Processing**: Vectorised, filtered to continental Europe (EPSG:3035), UK excluded

3.2 Overture Maps Foundation – 2024 Release

- **Source**: Overture Maps Foundation
- **URL**: `https://overturemaps.org/`
- **Release Date**: 2024-07-22 (alpha release)
- **Themes Used**:
  - Places (POIs): 2,000+ categories
  - Buildings: Footprints with attributes
  - Transportation: Road network (connectors and segments)
- **Access Method**: STAC (SpatioTemporal Asset Catalog)
- **License**: CDLA-Permissive-2.0 / ODbL (depending on source)
- **Processing**: Clipped to urban boundaries, transformed to EPSG:3035

3.3 Copernicus Urban Atlas 2018

- **Source**: European Environment Agency / Copernicus Land Monitoring Service
- **URL**: `https://land.copernicus.eu/local/urban-atlas`
- **Reference Year**: 2018
- **Resolution**: Minimum Mapping Unit (MMU) 0.25 ha for artificial surfaces
- **Classification**: 27 land cover/use classes
- **Coverage**: 788 Functional Urban Areas (FUAs) across EU
- **License**: Copernicus data policy (free and open access)
- **Processing**: Extracted urban blocks, green spaces; clipped to HDENS boundaries

3.4 Copernicus Street Tree Layer 2018

- **Source**: European Environment Agency / Copernicus Land Monitoring Service
- **URL**: `https://land.copernicus.eu/local/urban-atlas/street-tree-layer-stl-2018`
- **Reference Year**: 2018
- **Resolution**: 0.5m spatial resolution
- **Method**: Very High Resolution imagery interpretation
- **Coverage**: Street trees in Urban Atlas FUAs
- **License**: Copernicus data policy
- **Processing**: Clipped to HDENS boundaries; used for tree proximity calculations

3.5 Copernicus Building Height 2012

- **Source**: European Environment Agency / Copernicus Land Monitoring Service
- **URL**: `https://land.copernicus.eu/local/urban-atlas/building-height-2012`
- **Reference Year**: 2012
- **Resolution**: 10m × 10m raster
- **Method**: Digital Height Model (DHM) derived from stereo imagery
- **Units**: Metres above ground
- **Coverage**: Urban Atlas FUAs
- **License**: Copernicus data policy
- **Processing**: Sampled at building footprint centroids

3.6 Eurostat Census Grid 2021

- **Source**: Eurostat GEOSTAT project
- **URL**: `https://ec.europa.eu/eurostat/web/gisco/geodata/population-distribution/geostat`
- **Reference Year**: 2021 (Census round)
- **Resolution**: 1km × 1km grid (INSPIRE compliant)
- **Variables**: Population, employment, demographics, migration
- **Coverage**: EU27 + EFTA
- **License**: CC BY 4.0
- **Processing**: Interpolated to network nodes via grid cell centroids

4 Processing Parameters

4.1 Network Processing

- **CRS**: EPSG:3035 (ETRS89-LAEA Europe)
- **Decomposition**: 80m segments (≈1-minute walk at 80m/min)
- **Cleaning tolerance**: 8m (consolidation of degree-2 nodes)
- **Component threshold**: 100 nodes minimum
- **Level-aware processing**: Enabled (prevents incorrect bridge merging)
- **Edge classification**: Based on Overture connector/segment attributes

4.2 Centrality Computation

- **Method**: Segment-based (dual graph)
- **Thresholds**: [400, 800, 1200, 1600, 4800, 9600] metres
- **Walking times**: [5, 10, 15, 20, 60, 120] minutes
- **Walking speed**: 80 m/min (≈4.8 km/h)
- **Beta decay**: Exponential ($e^{-\beta \cdot d}$)
- **Min threshold weight**: 0.01832 (default cityseer value)

4.3 Accessibility Computation

- **Thresholds**: [400, 800, 1200, 1600] metres
- **Walking times**: [5, 10, 15, 20] minutes
- **Max assignment distance**: 100m (POI to nearest network edge)
- **Metrics**: Unweighted count, distance-weighted count, nearest distance
- **Weight function**: Exponential decay
- **Hill diversity**: q = [0, 1, 2] for richness, Shannon, Simpson

4.4 POI Saturation Assessment

- **Grid resolution**: 1km × 1km (Eurostat Census Grid)
- **Population scales**:
  - Local: 1km radius
  - Intermediate: 5km radius
  - Large: 10km radius
- **Model**: Random Forest Regressor (sklearn)
  - Trees: 100
  - Max depth: 20
  - Min samples split: 5
  - Random state: 42
- **Transformation**: Log-space ($\log(POI + 1)$ and $\log(pop + 1)$)
- **Z-score**: Standardised residuals from log-space predictions
- **Quadrant threshold**: Median standard deviation across categories

5 Software Dependencies

5.1 Core Packages

- **Python**: 3.12
- **cityseer**: 4.14.3 (network analysis, centrality, accessibility)
- **momepy**: 0.7.2 (morphometrics)
- **geopandas**: 1.0.1 (spatial data handling)
- **overturemaps**: 0.6.0 (Overture data access)
- **networkx**: 3.3 (graph data structures)
- **shapely**: 2.0.5 (geometric operations)
- **scikit-learn**: 1.5.1 (Random Forest regression)
- **rasterio**: 1.3.10 (raster data handling)
- **pyogrio**: 0.9.0 (fast vector I/O)

5.2 Environment Management

- **Package manager**: uv 0.4.15
- **Dependency specification**: pyproject.toml (PEP 621)
- **Reproducibility**: Pinned versions in pyproject.toml

6 Computational Requirements

6.1 Hardware

- **Minimum RAM**: 16 GB (32 GB recommended for largest cities)
- **Storage**: ≈500 GB for all 699 cities (complete pipeline outputs)
- **Processing time**: ≈2-5 hours per city (depending on size and complexity)

6.2 Parallelisation

- **City-level**: Fully independent (embarrassingly parallel)
- **Metric computation**: Single-threaded cityseer Rust backend
- **Recommended strategy**: Process multiple cities simultaneously

7 Data Quality Notes

7.1 Known Limitations

- **Temporal mismatch**: Source datasets span 2012-2024
  - Building heights: 2012
  - Urban Atlas: 2018
  - Census: 2021
  - Overture: 2024
- **POI completeness**: Variable across regions (see saturation assessment)
- **Building heights**: Not available for all FUAs; missing values set to estimated median
- **Network topology**: Minor inconsistencies at administrative boundaries

7.2 Quality Assurance

- **Network validation**: Automatic cleaning via cityseer (8m tolerance)
- **Topology checks**: Connected component analysis (min 100 nodes)
- **Attribute validation**: Range checks on computed metrics
- **Visual inspection**: Sample cities checked for geometric accuracy
- **Saturation assessment**: POI quality validated via multi-scale regression

8 Citation Information

8.1 Dataset Citation

When using this dataset, please cite:

> Simons, G. et al. (2025). SOAR: A Scalable, Open, Automated, and Reproducible Urban Data Model for the EU. *Data in Brief*, [volume]([issue]), [pages]. DOI: [to be assigned]

8.2 Software Citation

The cityseer package should be cited as:

> Simons, G. (2022). The cityseer Python package for pedestrian-scale network-based urban analysis. *Environment and Planning B: Urban Analytics and City Science*, 49(9), 2356–2361. `https://doi.org/10.1177/23998083221133827`

8.3 Source Data Citations

- **Overture Maps**: Overture Maps Foundation (2024). Overture Maps Data – 2024 Release. `https://overturemaps.org/`
- **HDENS-CLST**: European Commission, Joint Research Centre (2023). High Density Clusters – HDENS-CLST 2021. `https://ghsl.jrc.ec.europa.eu/`
- **Urban Atlas**: European Environment Agency (2020). Urban Atlas 2018. Copernicus Land Monitoring Service. `https://land.copernicus.eu/local/urban-atlas`
- **Census Grid**: Eurostat (2024). Census 2021 – Population Grid. `https://ec.europa.eu/eurostat/web/gisco/geodata/population-distribution/geostat`