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. Distance thresholds vary by metric type: centrality at 400m, 800m, 1200m, 1600m, 4800m; accessibility at 200m, 400m, 800m, 1200m, 1600m; morphology at 100m, 200m; green space proximity at 1600m; green area aggregation at 200m, 400m, 800m. Table 1 provides the complete attribute schema.

Table 1: Complete streets layer attribute schema. Suffix _DDD 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, DDD = 400, 800, 1200, 1600, 4800)			
`cc_beta_DDD`	Float	Beta-weighted closeness (gravity index with exponential decay)	cityseer
`cc_cycles_DDD`	Float	Cycle count within threshold	cityseer
`cc_density_DDD`	Float	Street segment density within threshold	cityseer
`cc_farness_DDD`	Float	Total network distance to reachable segments	cityseer
`cc_harmonic_DDD`	Float	Harmonic closeness (sum of inverse distances)	cityseer
`cc_hillier_DDD`	Float	Hillier-type closeness (node count / avg distance)	cityseer
`cc_betweenness_DDD`	Float	Segment betweenness centrality	cityseer
`cc_betweenness_beta_DDD`	Float	Beta-weighted betweenness centrality	cityseer
Accessibility – POI Counts (DDD = 200, 400, 800, 1200, 1600)			
`cc_accommodation_DDD_nw`	Integer	Unweighted count of accommodation POIs	cityseer
`cc_accommodation_DDD_wt`	Float	Distance-weighted count of accommodation POIs	cityseer
`cc_accommodation_nearest_max_DDD`	Float	Distance to nearest accommodation POI (m)	cityseer
`cc_active_life_DDD_nw/wt`	Float	Active life POI counts (unweighted/weighted)	cityseer
`cc_arts_and_entertainment_DDD_nw/wt`	Float	Arts and entertainment POI counts	cityseer
`cc_attractions_and_activities_DDD_nw/wt`	Float	Attractions and activities POI counts	cityseer
`cc_business_and_services_DDD_nw/wt`	Float	Business and services POI counts	cityseer
`cc_eat_and_drink_DDD_nw/wt`	Float	Eat and drink POI counts	cityseer
`cc_education_DDD_nw/wt`	Float	Education POI counts	cityseer
`cc_health_and_medical_DDD_nw/wt`	Float	Health and medical POI counts	cityseer
`cc_public_services_DDD_nw/wt`	Float	Public services POI counts	cityseer
`cc_religious_DDD_nw/wt`	Float	Religious POI counts	cityseer
`cc_retail_DDD_nw/wt`	Float	Retail POI counts	cityseer

Continued on next page

Table 1 – continued from previous page

Attribute	**Type**	**Description**	**Source/Method**
cc_{category}_nearest_max_DDD	Float	Distance to nearest POI in category (m)	cityseer
Infrastructure Accessibility (DDD = 200, 400, 800, 1200, 1600)			
cc_transport_DDD_nw/wt	Float	Transport stops/stations counts	cityseer
cc_street_furn_DDD_nw/wt	Float	Street furniture counts (benches, fountains)	cityseer
cc_parking_DDD_nw/wt	Float	Parking facility counts	cityseer
Mixed-Use Diversity Indices (DDD = 200, 400, 800, 1200, 1600)			
cc_hill_q0_DDD_nw	Float	Hill diversity (q=0, richness) unweighted	cityseer
cc_hill_q0_DDD_wt	Float	Hill diversity (q=0, richness) distance-weighted	cityseer
cc_hill_q1_DDD_nw	Float	Hill diversity (q=1, exp Shannon) unweighted	cityseer
cc_hill_q1_DDD_wt	Float	Hill diversity (q=1, exp Shannon) distance-weighted	cityseer
cc_hill_q2_DDD_nw	Float	Hill diversity (q=2, inv Simpson) unweighted	cityseer
cc_hill_q2_DDD_wt	Float	Hill diversity (q=2, inv Simpson) distance-weighted	cityseer
Building Morphology (DDD = 100, 200; stat = median, mad; suf = nw, wt)			
cc_area_{stat}_DDD_{suf}	Float	Building area statistics (m^2)	momepy
cc_perimeter_{stat}_DDD_{suf}	Float	Building perimeter statistics (m)	momepy
cc_compactness_{stat}_DDD_{suf}	Float	Circular compactness ($4\pi A/P^2$)	momepy
cc_orientation_{stat}_DDD_{suf}	Float	Orientation angle (degrees)	momepy
cc_mean_height_{stat}_DDD_{suf}	Float	Building height (m)	Copernicus DHM
cc_volume_{stat}_DDD_{suf}	Float	Building volume (m^3)	momepy
cc_floor_area_ratio_{stat}_DDD_{suf}	Float	Floor area ratio (3m floor height)	momepy
cc_form_factor_{stat}_DDD_{suf}	Float	Form factor ($A/(h \cdot P)$)	momepy
cc_corners_{stat}_DDD_{suf}	Float	Corner count	momepy
cc_shape_index_{stat}_DDD_{suf}	Float	Shape index	momepy
cc_shared_walls_{stat}_DDD_{suf}	Float	Shared wall length (m)	momepy
cc_fractal_dimension_{stat}_DDD_{suf}	Float	Fractal dimension	momepy
cc_building_DDD_nw/wt	Float	Building count (unweighted/weighted)	cityseer
Green Space Proximity (DDD = 200, 400, 800)			
cc_green_nearest_max_1600	Float	Distance to nearest green space (m)	Urban Atlas + cityseer
cc_trees_nearest_max_1600	Float	Distance to nearest tree canopy (m)	Street Tree Layer + cityseer
cc_green_area_sum_DDD_nw/wt	Float	Green space area within threshold (km^2)	Urban Atlas

Continued on next page

Table 1 – continued from previous page

Attribute	Type	Description	Source/Method
cc_trees_area_sum_DDD_nw/wt	Float	Tree canopy area within threshold (km^2)	Street Tree Layer
Demographics (interpolated from census grid)			
t	Float	Total population	Census 2021 + interpolation
density	Float	Population density (pop / land surface)	Census 2021
m, f	Float	Male / female population counts	Census 2021
m_%, f_%	Float	Male / female percentages	Census 2021
y_lt15, y_1564, y_ge65	Float	Age cohort counts (under 15, 15–64, 65+)	Census 2021
y_lt15_%, y_1564_%, y_ge65_%	Float	Age cohort percentages	Census 2021
emp	Float	Employment count	Census 2021
emp_%	Float	Employment percentage	Census 2021
nat, eu_oth, oth	Float	Nationality counts (national, EU other, non-EU)	Census 2021
nat_%, eu_oth_%, oth_%	Float	Nationality percentages	Census 2021
same, chg_in, chg_out	Float	Migration counts (same residence, in, out)	Census 2021
same_%, chg_in_%, chg_out_%	Float	Migration percentages	Census 2021
Block Characteristics (DDD = 100, 200; stat = median, mad; suf = nw, wt)			
cc_block_area_{stat}_DDD_{suf}	Float	Block area statistics (m^2)	Urban Atlas + momepy
cc_block_perimeter_{stat}_DDD_{suf}	Float	Block perimeter statistics (m)	Urban Atlas + momepy
cc_block_compactness_{stat}_DDD_{suf}	Float	Block circular compactness	Urban Atlas + momepy
cc_block_orientation_{stat}_DDD_{suf}	Float	Block orientation (degrees)	Urban Atlas + momepy
cc_block_covered_ratio_{stat}_DDD_{suf}	Float	Building coverage ratio	Urban Atlas + momepy
cc_block_DDD_nw/wt	Float	Block count (unweighted/weighted)	cityseer

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_and_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_and_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_and_entertainment	culture	cinema, theatre, concert_hall, opera_house
	entertainment	bowling, arcade, casino, nightlife_venue
attractions_and_activities	tourism	museum, gallery, monument, landmark
	attractions	zoo, aquarium, amusement_park, viewpoint
	nature	botanical_garden, nature_reserve, beach
religious	religious	church, mosque, synagogue, temple, monastery
Infrastructure Categories		
transport	transportation	bus_stop, tram_stop, metro_station, train_station, ferry_terminal
street_furn	amenity	bench, drinking_fountain, public_toilet, shelter
parking	parking	parking_lot, parking_garage, bicycle_parking

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 $\geq$1,500 inhabitants/km^2 and cumulative population $\geq$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] metres
- **Walking times**: [5, 10, 15, 20, 60] minutes
- **Walking speed**: 80 m/min (≈4.8 km/h)
- **Metrics**: beta, harmonic, hillier, farness, cycles, density, betweenness, betweenness_beta
- **Beta decay**: Exponential ($e^{-\beta \cdot d}$)
- **Min threshold weight**: 0.01832 (default cityseer value)

4.3 Accessibility Computation

- **Thresholds**: [200, 400, 800, 1200, 1600] metres
- **Walking times**: [2.5, 5, 10, 15, 20] minutes
- **Max assignment distance**: 100m (POI to nearest network edge)
- **Metrics**: Unweighted count, distance-weighted count, nearest distance
- **Categories**: 11 land-use (places) + 3 infrastructure
- **Weight function**: Exponential decay
- **Hill diversity**: q = [0, 1, 2] for richness, Shannon, Simpson

4.4 Morphology Computation

- **Thresholds**: [100, 200] metres
- **Statistics**: median, median absolute deviation (MAD)
- **Building metrics**: area, perimeter, compactness, orientation, corners, shape_index, fractal_dimension, shared_walls, volume, floor_area_ratio, form_factor
- **Block metrics**: area, perimeter, compactness, orientation, covered_ratio
- **Library**: momepy 0.7.2

4.5 Green Space Computation

- **Proximity threshold**: 1600 metres
- **Aggregation thresholds**: [200, 400, 800] metres
- **Sources**: Urban Atlas (green space classes 14100, 14200, 14400), Street Tree Layer

4.6 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, mixed-use, network-based aggregations)
- **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)

5.2 Environment Management

- **Package manager**: uv
- **Dependency specification**: pyproject.toml (PEP 621)
- **Reproducibility**: v1.0.0 release on Github repository

6 Computational Requirements

6.1 Hardware

- **Minimum RAM**: ≈32 GB recommended
- **Storage**: ≈500 GB recommended for processing all data and cities
- **Processing time**: ≈1 day for data preparation and ≈3 days for metric computation on M1 Macbook Pro

6.2 Parallelisation

- **Overture data**: Threads can be specifiied for speeding network cleaning
- **Metric computation**: Parallel processing supported via cityseer rust algorithms

7 Data Quality Notes

7.1 Known Limitations

- **Temporal mismatch**: Source datasets span 2012–2025
  - Building heights: 2012
  - Urban Atlas: 2018
  - Census: 2021
  - Overture: 2025
- **POI completeness**: Variable across regions (see saturation assessment)
- **Building heights**: Not available for all areas; missing values set to estimated median