Seven Exploratory Vignettes for the SOAR Urban Data Model: Illustrative Examples to Encourage Broader Urban Data Applications

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

[a] *University College London, United Kingdom*

Abstract

Large-scale urban datasets are often difficult to assess from outside: understanding what questions they can address requires substantial initial investment. This paper presents exploratory worked examples using the SOAR urban data model—a pan-European dataset covering 699 cities with over 100 metrics per street network node. These are not definitive research findings but illustrative vignettes showing types of analysis the data can support. We provide seven example templates demonstrating how workflows might leverage the data to explore data quality filtering, multi-scale analysis, access gap identification, predictive modelling, city benchmarking, classification, and site identification. Each vignette provides a reproducible workflow and suggests extensions. The purpose of these vignettes is to lower the barrier to entry for researchers considering whether SOAR suits their needs, not to make definitive claims about urban phenomena.

Keywords: urban data models, comparative urban analysis, walkability, data quality assessment, accessibility metrics, European cities, reproducible research, POI saturation

1. Introduction

Large-scale urban datasets can be difficult to evaluate from the outside. Understanding what data are contained, what questions can be addressed, and how workflows might be structured requires substantial effort before a researcher can determine whether a dataset suits their needs. This paper provides worked examples using the SOAR (Scalable, Open, Automated, and Reproducible) urban data model [1], illustrating the types of questions the data can support.

SOAR provides pre-computed metrics for 699 European urban centres, derived from Eurostat boundaries and demographics, Copernicus Urban Atlas land cover, and Overture Maps infrastructure data. The dataset includes over 100 metrics per street network node at multiple spatial scales (200–4,800m): network centrality, land-use accessibility, building morphology, green space proximity, and demographics.

This paper presents exploratory vignettes—worked examples that combine motivation, methodology, analysis, and interpretation. These are explicitly *not* definitive research findings. Each vignette is a starting point: a demonstration of what the data contain and how they might be used, not an exhaustive treatment of any

*Corresponding author
Email address: gareth.simons@ucl.ac.uk (Gareth Simons)

urban phenomenon. The vignettes are intentionally simple, using standard methods (correlations, Random Forests, clustering) rather than sophisticated causal inference or domain-specific theory. Researchers pursuing rigorous investigations should treat these as templates to adapt, not conclusions to cite. Each vignette includes:

- Research motivation
- SOAR metrics utilised
- Analytical workflow and code
- Results
- Possible extensions

The seven vignettes cover: data quality filtering (POI saturation assessment), multi-scale analysis (green space access), access gap identification (education, transit), predictive modelling (POI demand), benchmarking (15-minute cities), typology classification (urban morphology), and site selection (development opportunities).

The remainder of this paper is structured as follows: Section 2 reviews related work on urban data applications; Sections 3–9 present the seven vignettes; Section 10 discusses cross-cutting themes and limitations; Section 11 concludes.

2. Related Work

The development of large-scale urban datasets has been a major focus of recent urban science. Notable examples include the acquisition of 27,000 US street networks [2], the characterisation of 931 UK towns [3], and the creation of global feature-rich network datasets for 50 cities [4]. These efforts have been supported by the maturation of open-source toolsets for urban morphology and network analysis, such as `OSMnx` [2], `momepy` [5], and `cityseer` [6].

The SOAR data model builds on these foundations by providing pan-European coverage with a consistent set of over 100 metrics per street network node. This enables comparative urban analysis at a continental scale, addressing questions of walkability, accessibility, and urban form across diverse national contexts. The vignettes presented in this paper demonstrate how these data can be applied to common urban research tasks, from data quality assessment to site selection.

3. Vignette 1: Data Quality Filtering

3.1. Motivation

POI completeness varies geographically, with some regions exhibiting systematic undersaturation. This vignette applies multi-scale regression of POI counts against population densities to identify cities where crowdsourced data may be too sparse for reliable analysis, allowing researchers to filter or weight observations accordingly.

3.2. SOAR Metrics Utilised

- **POI counts**: 11 land-use categories (accommodation, active life, arts & entertainment, attractions, business services, eat & drink, education, health & medical, public services, religious, retail)
- **Census demographics**: Population counts at 1 km^2 grid resolution
- **Multi-scale neighbourhoods**: Local (2 km), intermediate (5 km), and large (10 km) radii

3.3. Methodology

We develop a grid-based multi-scale regression approach to assess POI data saturation across cities, comparing observed POI densities against population-based expectations to identify undersaturated areas that may indicate data incompleteness. This method provides a quantitative foundation for evaluating data quality prior to comparative urban analysis.

3.3.1. Multi-Scale Regression Workflow

The saturation assessment workflow (`paper_research/code/eg1_data_quality/`) operates at the 1 km^2 census grid level, enabling fine-grained spatial analysis:

1. **Grid-level aggregation**: POI counts are computed within each census grid cell. Multi-scale population neighborhoods are calculated at local, intermediate, and large radii to capture hierarchical catchment effects.

2. **Random Forest regression**: For each land-use category k, a Random Forest model is fitted in log-space:

$$\log(\text{POI}_k + 1) = f(\log(\text{pop}_{\text{local}}), \log(\text{pop}_{\text{intermediate}}), \log(\text{pop}_{\text{large}})) + \epsilon \tag{1}$$

   Log transformation linearizes the power-law relationship between population and POI counts (POI $\propto$ pop^{β}), yielding more normally distributed residuals suitable for z-score computation.

3. **Z-score computation**: Standardized residuals quantify deviation from expected POI counts. Negative z-scores indicate undersaturation (fewer POIs than expected); positive z-scores indicate saturation.

4. **City-level aggregation**: Grid z-scores are aggregated per city, computing mean (overall saturation level) and standard deviation (spatial variability within city).

5. **Quadrant classification**: Cities are classified by mean z-score $\times$ variability into four quadrants: consistently undersaturated, variable undersaturated, consistently saturated, and variable saturated.

3.3.2. Quadrant Interpretation

The quadrant classification provides actionable guidance for data usage:

- **Consistently Undersaturated** (low mean, low std): Systematic data gaps; use with caution across all analyses

- **Variable Undersaturated** (low mean, high std): Partial coverage; some grid cells may be reliable
- **Consistently Saturated** (high mean, low std): Complete coverage; suitable for all analyses
- **Variable Saturated** (high mean, high std): Good overall coverage with spatial heterogeneity

3.4. Results

Analysis of 699 European urban centres reveals a core-periphery pattern in POI data saturation. Central and Western European cities (Germany, Netherlands, France, Belgium) achieve mean z-scores near zero with low spatial variability, indicating reliable data. Peripheral European regions show systematic undersaturation: Spanish cities (particularly Madrid satellites) average -0.6 to -1.1, with similar patterns in Romania, Bulgaria, Poland, and southern Italy.

This pattern likely reflects differential OpenStreetMap contributor activity, varying commercial formalisation practices, and regional differences in POI aggregator coverage. The effect is pronounced for business services and retail (R^2=0.73, 0.70), while accommodation shows weakest predictability (R^2=0.56), suggesting tourism infrastructure follows different spatial logic.

Table 1 summarises Random Forest model performance by POI category. R^2 values range from 0.56 (accommodation) to 0.73 (business services), with local population scale consistently the strongest predictor for everyday amenities (retail, eat_and_drink, health_and_medical) while intermediate-scale population better predicts destination categories (attractions_and_activities).

Table 1: Random Forest regression performance by POI category. Local, intermediate, and large columns show relative feature importance for each population scale.

Category	**R^2**	**Local**	**Intermed.**	**Large**
Business And Services	0.73	0.76	0.14	0.10
Education	0.73	0.72	0.16	0.12
Eat And Drink	0.72	0.72	0.15	0.12
Retail	0.70	0.75	0.14	0.12
Health And Medical	0.69	0.72	0.14	0.14
Public Services	0.69	0.65	0.20	0.15
Active Life	0.65	0.64	0.21	0.15
Arts And Entertainment	0.63	0.48	0.33	0.19
Attractions And Activities	0.60	0.28	0.52	0.21
Religious	0.59	0.56	0.23	0.21
Accommodation	0.56	0.41	0.33	0.26

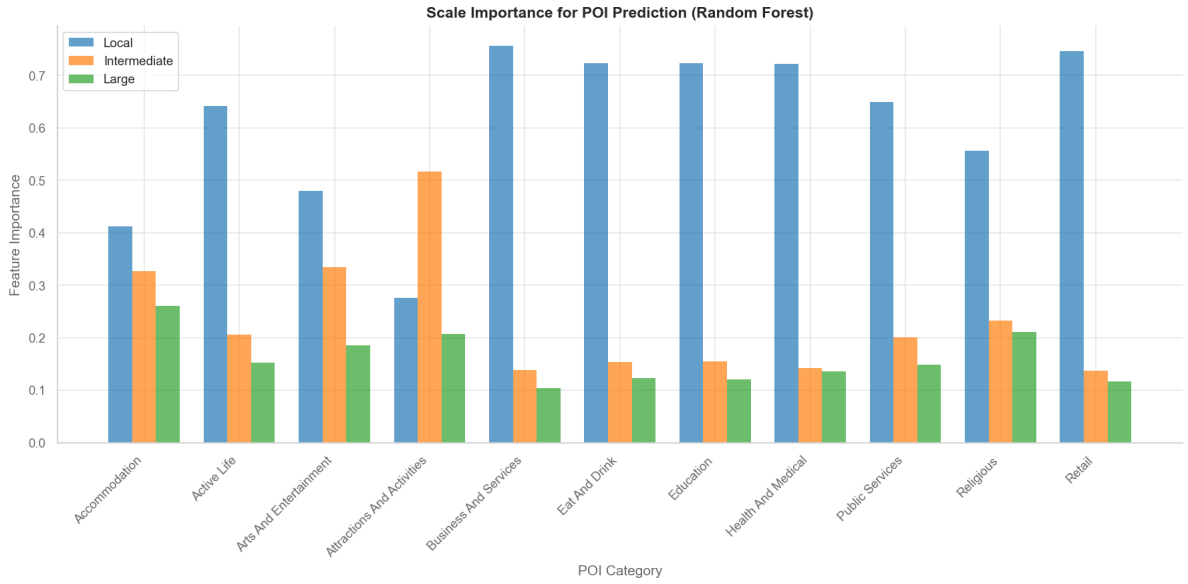

Figure 1: Feature importance showing which population scale (local, intermediate, large) best predicts POI distribution for each category.

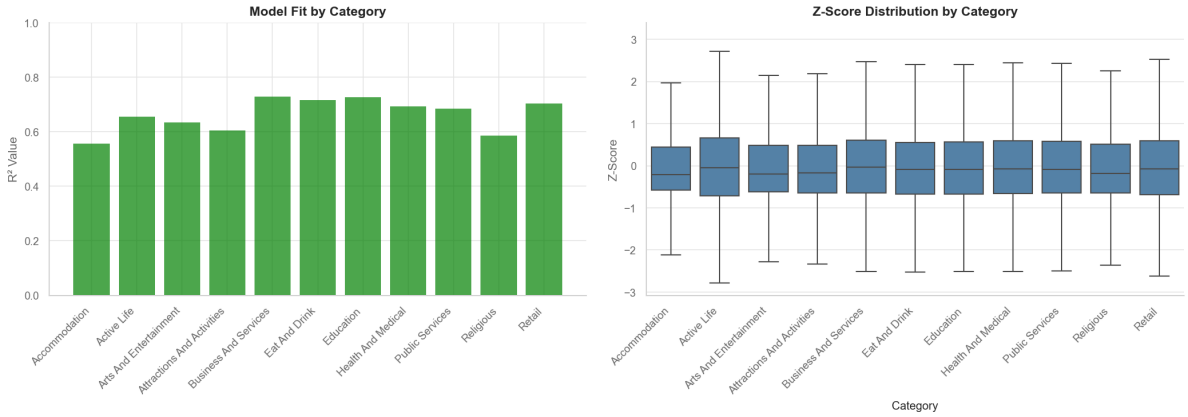

Figure 2: Exploratory data analysis. Left: Random Forest model fit (R^2) by POI category. Right: distribution of z-scores across grid cells per category.

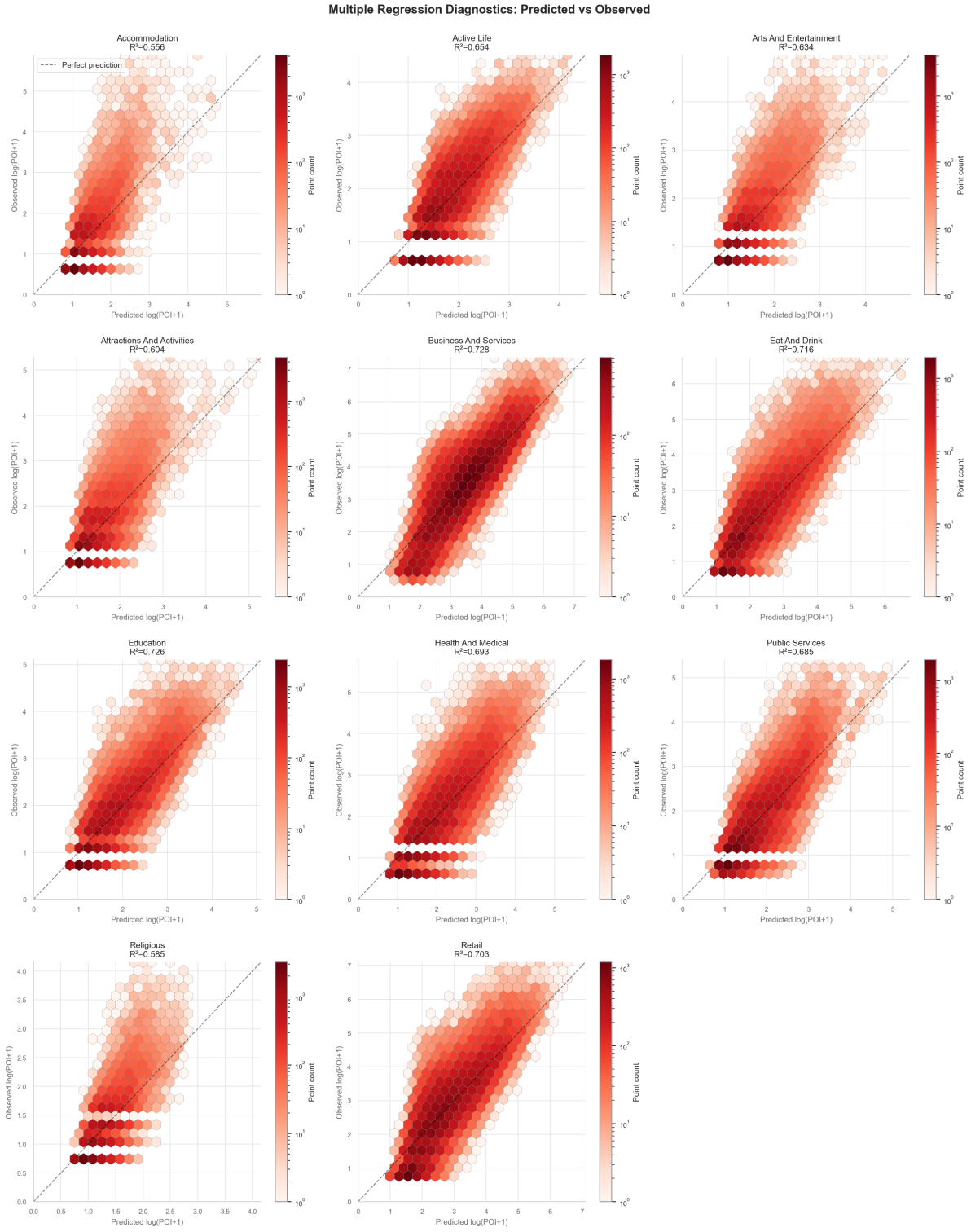

Figure 3: Regression diagnostics: predicted vs. observed POI counts (log scale) for each category.

Figure 4: City quadrant analysis. X-axis: mean z-score (negative = undersaturation). Y-axis: standard deviation (within-city variability). Quadrant colours: red = consistently undersaturated; green = consistently saturated; orange = variable undersaturated; blue = variable saturated.

3.5. Implications

Researchers comparing POI-derived metrics across European cities should account for systematic data quality variation. Options include: (1) restricting analyses to consistently saturated cities; (2) stratifying by saturation quadrant; or (3) applying z-score corrections in undersaturated regions.

3.6. Extensions

Potential directions: temporal trends in POI completeness; category-specific quality metrics; validation against municipal records; correlations between data quality and urban characteristics; saturation vectors as city-level features.

3.7. Reproducibility

Code, outputs, and documentation: `https://github.com/[repo]/paper_research/code/eg1_data_quality/`

4. Vignette 2: Multi-Scale Analysis

4.1. Motivation

Relationships observed within cities may differ from those observed between cities. SOAR enables both within-city analysis and city-level aggregations for cross-city comparison, and the two perspectives can yield different—sometimes opposing—conclusions. This vignette explores the topic in the context of access to green spaces, examining within-city correlations between population density and green space proximity.

4.2. SOAR Metrics Utilised

- **Green space accessibility**: Network distance to nearest green block (1,600m catchment)
- **Tree canopy accessibility**: Network distance to nearest tree canopy (1,600m catchment)
- **Population density**: Persons per km^2 (interpolated from Eurostat 1km grid)

4.3. Methodology

For each city with $\geq$100 street network nodes, we compute Spearman rank correlations between population density and distance to green space/tree canopy. Negative correlations indicate compact urban cores with proximate green access ("dense-and-green"), while positive correlations suggest peripheral green amenities with undersupplied centres ("dense-but-grey"). Results are visualised as diverging bar charts sorted by correlation strength, with cities categorised by the direction and magnitude of their density-green relationship.

4.4. Results

Analysis of 672 cities across 16.6 million street network nodes suggests a consistent within-city pattern for green blocks alongside contrasting behavior for tree canopy:

Green space (parks): 666 cities (99%) exhibit positive correlations, where denser areas face longer walks to parks. Median distance is 93.5m, with 91.0% of nodes within a 5-minute walk (400m). The strongest positive correlation (Verviers, Belgium: $\rho = 0.81$) exemplifies peripheral park placement, while rare negative outliers like Meiderich/Beeck, Germany ($\rho = -0.07$) and Płock, Poland ($\rho = -0.05$) demonstrate integrated green infrastructure in high-density zones.

Tree canopy: 569 cities (85%) show negative correlations, indicating that denser neighbourhoods have *better* tree canopy access. Median distance is 72.2m, with 91.2% within 400m. Strong negative correlations (e.g., Valdemoro, Spain: $\rho = -0.67$; Coslada, Spain: $\rho = -0.60$) suggest street tree programmes concentrated in urban cores, likely reflecting municipal maintenance priorities and sidewalk infrastructure availability.

4.5. Discussion

The data show consistent within-city patterns: in most cities, denser areas are farther from parks but closer to street trees. There is no systematic cross-city pattern—dense cities are not inherently worse for green access than sparse cities.

These are descriptive observations, not causal claims. The patterns could reflect land economics (parks in cheaper peripheries), planning decisions (street trees prioritised in pedestrian areas), or other factors. More rigorous investigation would require historical analysis, policy review, or quasi-experimental designs.

4.6. Extensions

Potential directions: green space quality metrics; temporal analysis of densification; behavioural validation; green space typology effects; policy mechanism studies.

4.7. Reproducibility

Code, outputs, and documentation: `https://github.com/[repo]/paper_research/code/eg2_multiscale/`

5. Vignette 3: Access Gap Identification

5.1. Motivation

Distance-to-nearest metrics can reveal locations where distances to amenities or services are greater than average or exceed targeted thresholds. This vignette explores where distances to education and transport are greater than typical, helping to highlight areas of potential disadvantage warranting further investigation.

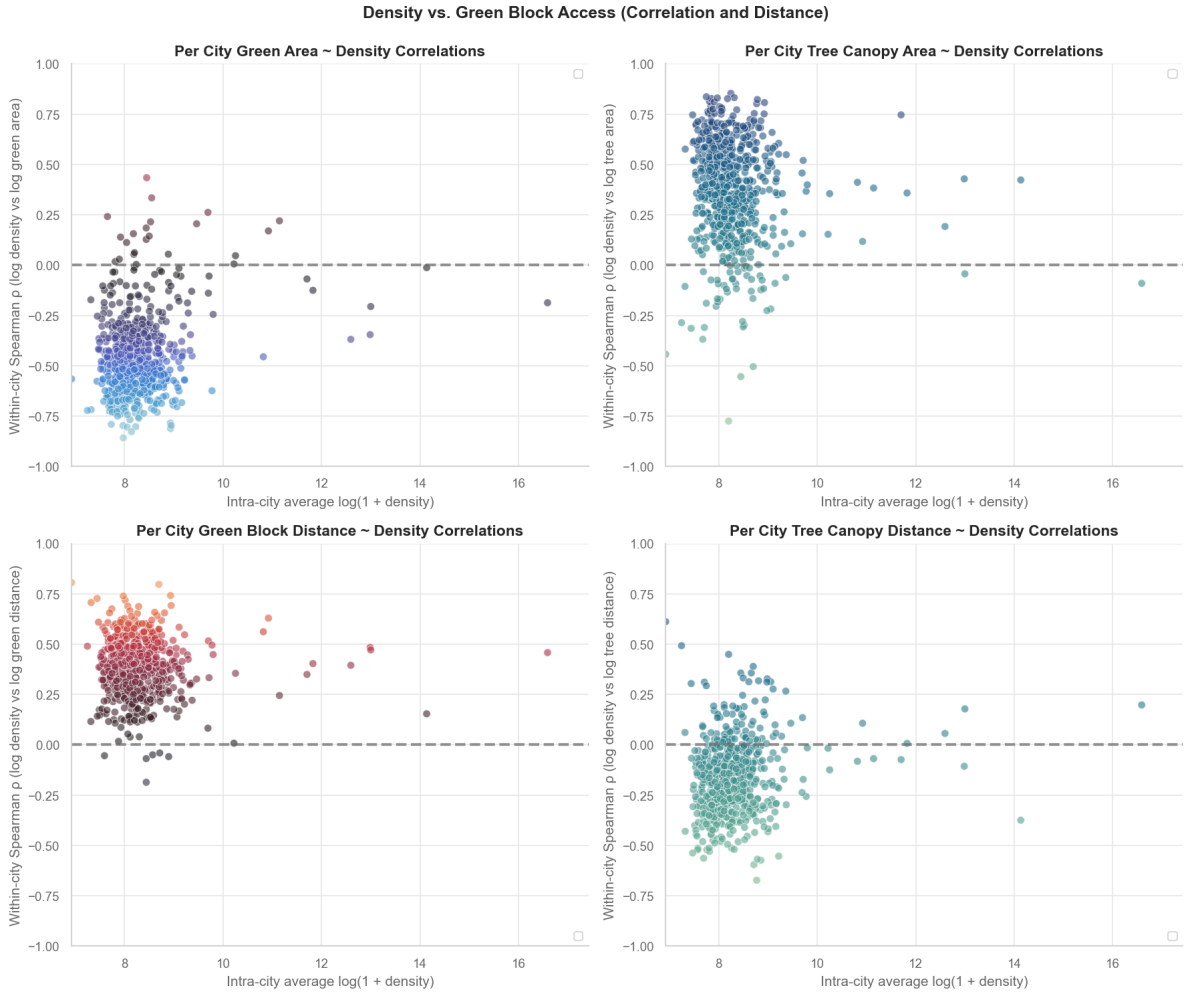

Figure 5: **Green space accessibility and tree canopy versus population density.** 2×2 grid comparing distance metrics (top row) and correlation analysis (bottom row) across 491 European cities. Left column: green blocks (parks). Right column: tree canopy. Points colored by Spearman correlation strength (blue=negative, red=positive). Top panels show no systematic relationship between city-level density and mean green distance; bottom panels confirm the absence of cross-city patterns for density-access correlations.

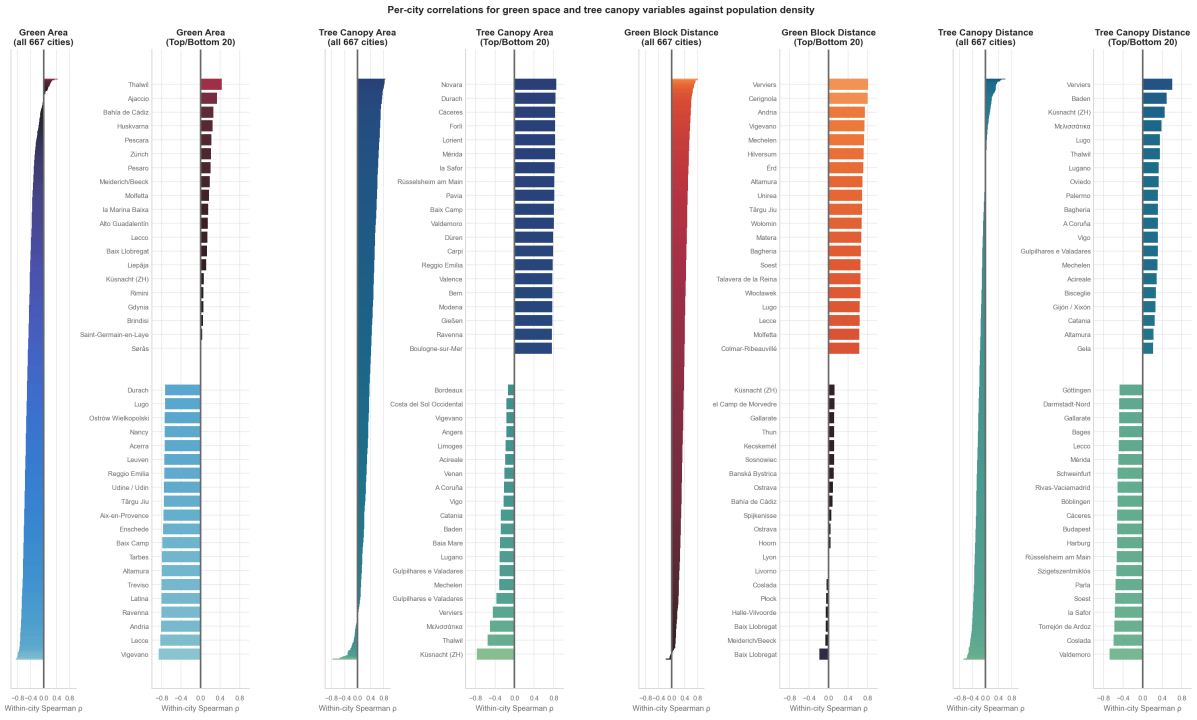

Figure 6: **Per-city correlation patterns for green space accessibility.** Diverging bar chart ranking cities by density-green correlation. For green blocks: only 4 cities show negative correlations (dense neighborhoods closer to parks); 487 cities show positive correlations (parks in peripheries). For tree canopy: 478 cities show negative correlations (street trees in urban cores); only 13 show positive correlations.

5.2. SOAR Metrics Utilised

- `cc_education_nearest_max_1600`: Network distance to nearest education POI (m)
- `cc_transportation_nearest_max_1600`: Network distance to nearest transport stop (m)
- `cc_beta_800`: Local network centrality (demand proxy)
- `density`: Population density (demand proxy)

5.3. Methodology

We conduct two complementary analyses:

Education access: For cities with `Consistently Saturated` education POI coverage (from Vignette 1), we compute mean and median network distances to the nearest school, along with the proportion of nodes within 400m and 800m walking distance. To capture spatial equity, we calculate the P75/P25 ratio and the percentage of nodes with access worse than twice the city mean.

Transport gaps: For cities with reliable transport POI coverage, we identify locations where high demand (based on network centrality and population density) coincides with poor transport supply (long distances to nearest stop). Gap areas are classified using percentile thresholds: high-demand nodes (top 30%) with low transport supply (bottom 30%) are flagged as gap areas; those with critically low supply (bottom 15%) are flagged as critical gaps.

5.4. Results

Education access varies substantially across Europe. Cities like Küsnacht (CH) and Płock (PL) have over 65% of nodes within a 5-minute walk of a school, while cities like Como (IT) and Iserlohn (DE) have mean distances exceeding 600m. Table 2 highlights the top and bottom performers.

Table 2: Education access: top and bottom cities by mean distance and % within 400m.

City	Country	Mean Dist. (m)	% within 400m
Küsnacht (ZH)	CH	328	69.8
Płock	PL	376	69.4
Hoorn	NL	388	64.3
A Coruña	ES	395	61.8
Leiden	NL	396	65.0
...			
Aschaffenburg	DE	583	44.1
Lüdenscheid	DE	584	41.2
Sosnowiec	PL	593	39.9
Wołomin	PL	596	40.0
Schweinfurt	DE	612	35.5

Equity is not guaranteed by abundance. Even in cities with good average access, pockets of disadvantage persist. The P75/P25 ratio ranges from approximately 2.4 to over 6, and in the least equitable cities, a substantial fraction of nodes are severely underserved (Table 3).

Table 3: Most and least equitable cities by P75/P25 ratio.

City	Country	P75/P25 Ratio	% Severely Underserved
Nieuwegein	NL	2.4	9.8
Almere	NL	2.5	10.9
Almelo	NL	2.5	10.1
...			
Ξάνθη	GR	6.2	18.5
Toledo	ES	6.0	19.4
Hoya de Huesca / Plana de Uesca	ES	5.7	18.3

5.5. Discussion

Educational access varies substantially across Europe, with mean distances ranging from approximately 330m to over 640m across the analysed cities. The P75/P25 ratio—ranging from approximately 2.4 to over 6—distinguishes cities with equitable distribution from those with concentrated provision. Notably, cities with good average access may still have substantial pockets of disadvantage: even well-performing cities show 8–20% of nodes classified as severely underserved. This highlights why average access metrics can mask substantial within-city disparities.

The methodology outlined for transport gap analysis demonstrates how demand-supply mismatches can be identified by combining centrality (as a proxy for activity potential) with accessibility metrics. Locations where high network centrality and population density coincide with poor transport access represent candidates for further investigation. These are descriptive findings that could inform targeted infrastructure planning, not causal claims about infrastructure adequacy.

5.6. Extensions

Potential directions: school capacity and enrolment data; socioeconomic correlates of underserved areas; scenario modelling for new transport stops; temporal trend analysis; case studies of identified gap areas; validation against actual travel behaviour.

5.7. Reproducibility

Code, outputs, and documentation: `https://github.com/[repo]/paper_research/code/eg3_access_gaps/`

6. Vignette 4: Predictive Modelling

6.1. Motivation

Large-scale datasets from multiple cities can enable training of generalisable models. This vignette uses network centrality and population density to predict levels of eating and drinking establishments as well as business and services intensities. The consistent feature set across cities can support transfer learning or pooled models.

6.2. SOAR Metrics Utilised

- **Network centrality**: Closeness centrality at 400m, 800m, 1,200m, 1,600m, and 4,800m radii
- **POI counts**: Eat & drink establishments (400m); Business & services establishments (400m)
- **Census variables**: Population density, age structure (under 15, 15–64, 65+), and employment ratio (interpolated from Eurostat 1km grid)
- **Saturation classification**: From Vignette 1 (cities classified as `Consistently Saturated`)

6.3. Methodology

We develop an Extra Trees regression approach to predict node-level POI counts based on multi-scale network centrality and census demographics. Analysis is restricted to cities with `Consistently Saturated` POI coverage for both eat & drink and business & services categories (as identified in Vignette 1), yielding 27 cities and 694,527 street network nodes.

6.3.1. Model Training Workflow

The workflow (`paper_research/code/eg4_prediction/`) operates at the street network node level:

1. **Data preparation**: Extract closeness centrality metrics at five spatial scales (400–4,800m) and POI counts for eat & drink and business & services categories. Filter to cities with consistent saturation to avoid training on incomplete data.

2. **Feature engineering**: Combine 5 centrality features with 5 census features (density, age groups, employment). Log-transform all features and targets: $\log(x + 1)$.

3. **Train-test split**: Randomly split nodes into 90% training and 10% testing sets, stratified by city to ensure geographic representation.

4. **Extra Trees training**: Fit separate models for eat & drink and business & services. Hyperparameters: 100 estimators, maximum depth of 20.

5. **Per-city evaluation**: Compute R^2, MAE, and RMSE separately for each city to assess how well the centrality-amenity relationship generalises across urban contexts.

6.4. Results

Extra Trees models achieve strong predictive performance on held-out test data: $R^2 = 0.731$ for both eat & drink and business & services. Per-city R^2 values reveal how consistently the centrality-amenity relationship holds across different urban forms.

Eat & drink: Median city $R^2 = 0.709$, with 96.3% of cities achieving $R^2 > 0.5$. Italian and Spanish cities dominate the best-predicted list (Table 4), suggesting that network structure strongly determines hospitality location in these contexts. Heerlen ($R^2 = 0.44$) shows poorest fit, potentially indicating amenity distributions driven by factors beyond network accessibility.

Business & services: Median city $R^2 = 0.721$, with 100% of cities exceeding $R^2 > 0.5$. The pattern mirrors eat & drink, with Italian cities showing strongest network-amenity alignment (Table 5).

Table 4: Top 10 cities by R^2 for Eat & Drink (400m).

City	R^2	MAE	RMSE	Nodes
Bari	0.857	0.430	0.584	14,855
la Safor	0.821	0.425	0.603	5,435
Ragusa	0.819	0.375	0.512	6,693
la Plana Alta	0.816	0.448	0.629	10,265
Alessandria	0.809	0.458	0.590	5,417
...				
Heerlen	0.440	0.488	0.611	33,426
Pordenone / Pordenon	0.585	0.497	0.655	10,823
Gallarate	0.586	0.479	0.606	17,720
Bergamo	0.590	0.496	0.638	41,833

Table 5: Top 10 cities by R^2 for Business & Services (400m).

City	R^2	MAE	RMSE	Nodes
Ragusa	0.864	0.406	0.571	6,693
Cremona	0.816	0.498	0.643	6,138
Bari	0.809	0.549	0.725	14,855
la Safor	0.804	0.550	0.759	5,435
Alessandria	0.798	0.536	0.712	5,417
...				
Modena	0.564	0.679	0.892	14,782
Heerlen	0.594	0.567	0.726	33,426
Pordenone / Pordenon	0.622	0.601	0.771	10,823
Prato	0.637	0.639	0.796	20,100

Feature importance reveals that intermediate-scale closeness centrality (1,200–1,600m) dominates predictions for both categories (Table 6), consistent with pedestrian catchment theory—amenities locate where they can serve walkable neighbourhoods rather than immediate adjacency (400m) or regional accessibility (9,600m). Census features contribute substantially, with employment ratio and population density ranking among the top predictors.

Table 6: Feature importance for Extra Trees regression models.

Eat & Drink		Business & Services	
Feature	Importance	Feature	Importance
Cc Beta 1200	0.173	Cc Beta 1200	0.214
Cc Beta 1600	0.152	Cc Beta 1600	0.171
Y 1564	0.126	Cc Beta 800	0.145
Cc Beta 800	0.117	Y 1564	0.093
Y Ge65	0.092	Emp	0.068
Density	0.083	Y Ge65	0.067
Emp	0.079	Y Lt15	0.067
Y Lt15	0.072	Density	0.057
Cc Beta 4800	0.048	Cc Beta 4800	0.046
Cc Beta 400	0.032	Cc Beta 400	0.040

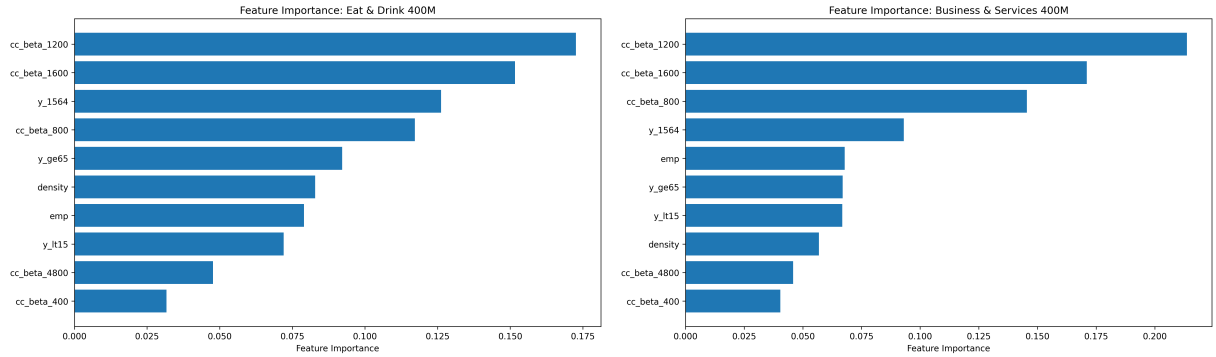

Figure 7: **Feature importance for amenity prediction models.** Left: Eat & drink. Right: Business & services. Intermediate-scale closeness centrality (1,200–1,600m) dominates both models, consistent with pedestrian catchment theory.

6.5. Discussion

Models achieve $R^2 > 0.7$ on held-out data, indicating that network centrality and census features correlate with amenity counts. Intermediate-scale centrality (1,200–1,600m) dominates feature importance.

These are correlational findings. The models do not establish that centrality causes commercial location; both could be driven by underlying factors (land values, zoning, historical development). Per-city variation in model performance suggests the relationship differs across urban contexts.

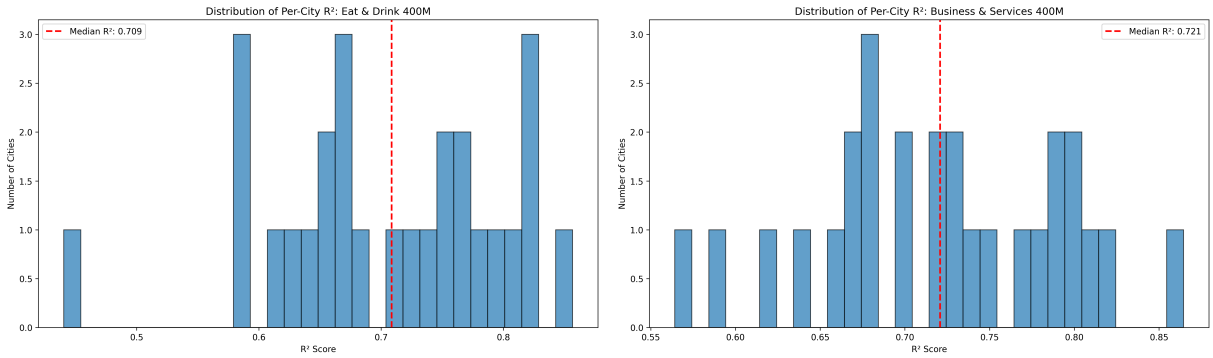

Figure 8: **Distribution of per-city R^2 scores.** Left: Eat & drink. Right: Business & services. Most cities cluster at high R^2 values, with a few outliers showing weaker network-amenity relationships.

6.6. Extensions

Potential directions: additional explanatory variables (land values, zoning); residual analysis to identify underserved areas; temporal stability; cross-category comparisons; city-specific models.

6.7. Reproducibility

Code, outputs, and documentation: `https://github.com/[repo]/paper_research/code/eg4_prediction/`

7. Vignette 5: Benchmarking

7.1. Motivation

Cities can be ranked on standardised metrics enabling comparative assessment against peer cities or policy targets. This vignette ranks cities by walkable access to amenities and services and identifies those with best overall access to all land-use categories, operationalising the "15-minute city" concept using SOAR's pre-computed POI distances.

7.2. SOAR Metrics Utilised

- **POI network distances**: `cc_{category}_nearest_max_1600` for 10 essential categories
- **POI categories assessed**: Active life, arts & entertainment, attractions & activities, business & services, eat & drink, education, health & medical, public services, religious, retail (accommodation excluded as non-essential for daily access)
- **Saturation classification**: From Vignette 1 (cities with `Consistently Saturated` or `Variable Saturated` combined POI coverage)

7.3. Methodology

We develop a node-level completeness scoring approach to benchmark cities against the 15-minute city ideal. Analysis is restricted to cities with reliable POI coverage across multiple categories (as identified in Vignette 1's between-category quadrant classification).

7.3.1. Completeness Scoring Workflow

The workflow (`paper_research/code/eg5_benchmarking/`) operates at the street network node level:

1. **City filtering**: Select cities classified as `Consistently Saturated` or `Variable Saturated` in the between-category quadrant analysis (Vignette 1), ensuring reliable POI data across multiple service types.

2. **Distance extraction**: For each node, extract network distances to nearest POI in each of 10 categories using SOAR's pre-computed `cc_{category}_nearest_max_1600` metrics.

3. **Per-node completeness**: Count how many of the 10 categories are accessible within 1,200m (approximately 15 minutes at 80m/min walking speed). A node with "full access" reaches all 10 categories within this threshold.

4. **City-level aggregation**: Compute the percentage of nodes with full access (all 10 categories within 1,200m), mean completeness score (average number of accessible categories divided by 10), and per-category access rates.

5. **Bottleneck identification**: Identify which POI categories most frequently limit full access, revealing systematic infrastructure gaps.

7.3.2. Threshold Rationale

The 1,200m threshold operationalises a 15-minute walk assuming approximately 80m/min walking speed, consistent with pedestrian planning standards. This threshold captures a realistic daily walking catchment while being stringent enough to distinguish genuinely walkable neighbourhoods from car-dependent areas.

7.4. Results

Analysis of cities with reliable POI coverage suggests substantial variation in 15-minute city completeness across European urban centres.

Most cities achieve substantial 15-minute completeness, though universal coverage remains elusive. The median city has 78.3% of nodes with access to all 10 POI categories within 1,200m, with a range from 48.5% to 94.5%. While this suggests many European cities have strong walkable infrastructure, the gap between median and best performers indicates room for improvement—even cities with good overall accessibility may have neighbourhoods lacking coverage across all service types.

Top-performing cities in this analysis demonstrate that compact urban form and mixed-use development can deliver near-complete 15-minute access. Table 7 shows cities with highest percentages of fully-accessible nodes. These cities share characteristics of pedestrian-oriented development, fine-grained land-use mixing, and comprehensive local service provision.

Table 7: Top 10 cities by 15-minute city completeness.

City	Country	% Full Access	Mean Completeness
Venezia	IT	94.5	0.991
Bahía de Cádiz	ES	93.5	0.990
Siracusa	IT	92.9	0.983
Trapani	IT	90.7	0.983
Amersfoort	NL	90.1	0.984
Pesaro	IT	89.2	0.980
Verona	IT	88.6	0.979
Gouda	NL	88.5	0.978
Firenze	IT	88.5	0.978
Ancona	IT	88.1	0.977

Bottom-performing cities in this analysis exhibit either sprawling urban form, car-oriented development patterns, or systematic gaps in specific service categories. Table 8 highlights cities with lowest completeness scores.

Table 8: Bottom 10 cities by 15-minute city completeness.

City	Country	% Full Access	Mean Completeness
Gulpilhares e Valadares	PT	48.5	0.918
Gulpilhares e Valadares	PT	57.3	0.932
Gulpilhares e Valadares	PT	58.6	0.939
Como	IT	59.1	0.929
Hradec Králové	CZ	61.4	0.950
None	None	61.9	0.943
Roosendaal	NL	62.0	0.952
Roanne	FR	62.1	0.950
Thalheim bei Wels	AT	62.2	0.939
Toulon	FR	62.6	0.943

National differences are also evident. Table 9 aggregates city scores by country, showing that Southern European countries (Spain, Romania, Croatia, Italy) tend to have higher proportions of fully accessible nodes compared to Nordic countries (Sweden, Denmark, Norway, Finland), likely reflecting differences in urban density and zoning traditions.

Bottleneck categories reveal which services most frequently limit full 15-minute access. Table 10 ranks POI categories by mean access rate, identifying systematic gaps that planners could target for intervention.

Table 9: Country rankings by mean 15-minute city completeness score.

Country	Cities	Mean % Full Access	Median % Full Access
ES	11	82.9	81.9
RO	7	82.7	83.0
HR	5	82.4	82.5
IT	62	81.7	82.8
GR	4	79.9	81.1
...			
SE	7	72.4	71.5
DK	4	71.9	70.6
NO	3	70.8	72.6
PT	8	70.5	77.0
FI	4	66.9	65.4

Table 10: POI categories ranked by mean 15-minute access rate (bottleneck analysis).

Category	Mean Access Rate (%)
Religious	87.4
Arts and Entertainment	92.5
Attractions and Activities	94.2
Education	97.3
Health and Medical	97.7
Public Services	97.9
Active Life	98.2
Eat and Drink	98.8
Retail	99.5
Business and Services	99.9

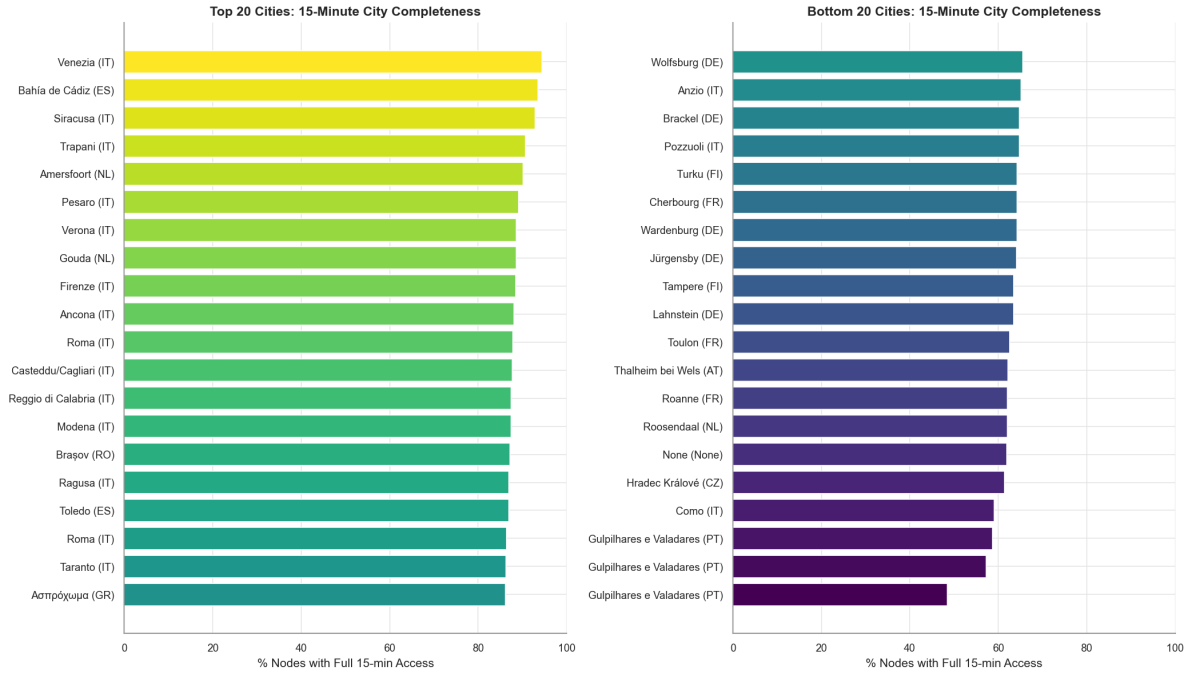

Figure 9: **15-minute city completeness ranking.** Left: Top 20 cities with highest proportion of nodes achieving full 15-minute access. Right: Bottom 20 cities. Colour intensity reflects completeness score.

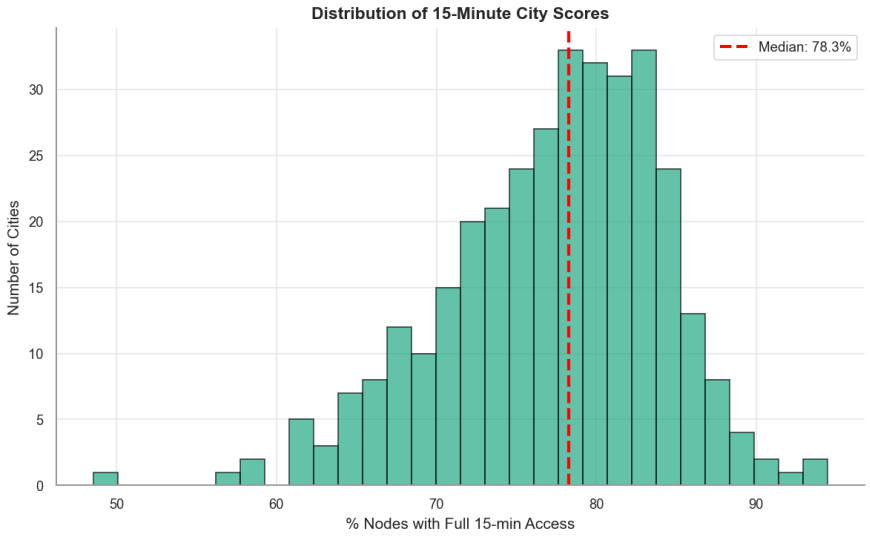

Figure 10: **Distribution of 15-minute city scores across European cities.** Histogram showing the percentage of nodes with full access to all 10 POI categories. Red dashed line indicates median.

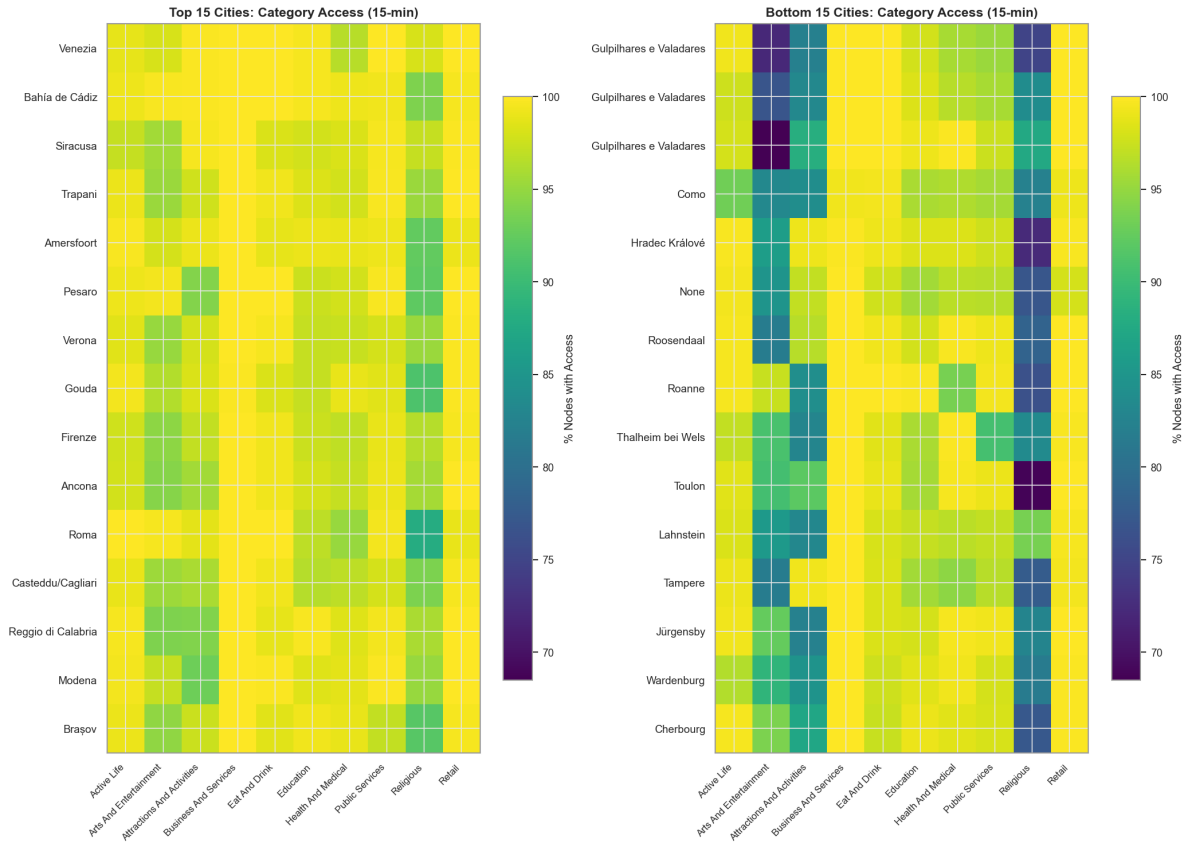

Figure 11: **Per-category access rates for top and bottom cities.** Heatmap showing which POI categories are most/least accessible within 1,200m. Columns represent categories; rows represent cities. Darker colours indicate lower access rates, revealing category-specific bottlenecks.

7.5. Discussion

The analysis reveals that full 15-minute access is typically limited by one or two categories rather than general deficits across all services. Religious facilities (87.4% mean access) and arts & entertainment venues (92.5%) consistently emerge as bottleneck categories, while business & services (99.9%) and retail (99.5%) achieve near-universal coverage. This pattern suggests that commercial activities locate to maximise accessibility, while specialised cultural and religious facilities serve larger catchments.

Top-performing cities (Venezia at 94.5%, Bahía de Cádiz at 93.5%) share characteristics of compact historic cores with fine-grained land-use mixing. Italian cities dominate the top rankings, appearing 7 times in the top 10. Bottom performers tend to be smaller cities or suburban agglomerations where dispersed development patterns limit walkable service provision.

National patterns are evident: Southern European countries (Italy, Spain, Croatia) show higher median completeness than Nordic countries (Sweden, Denmark, Finland), likely reflecting historical urban density, zoning traditions, and car-oriented versus pedestrian-oriented development. However, these patterns should be interpreted cautiously—POI data completeness (Vignette 1) may also vary systematically by region.

This metric is sensitive to category definitions and distance thresholds; different operationalisations would yield different rankings. The 1,200m threshold operationalises one interpretation of "15-minute city," but actual travel times depend on terrain, pedestrian infrastructure quality, and individual mobility.

7.6. Extensions

Potential directions: category weighting by use frequency or essentiality; actual routing travel times accounting for pedestrian infrastructure; opening hours analysis to capture temporal accessibility; cycling vs walking comparisons; socioeconomic correlates of access gaps; scenario modelling for targeted infrastructure investment.

7.7. Reproducibility

Code, outputs, and documentation: `https://github.com/[repo]/paper_research/code/eg5_benchmarking/`

8. Vignette 6: Typology Classification

8.1. Motivation

Clustering algorithms applied to street-level features can identify recurring neighbourhood types that transcend administrative boundaries, revealing morphological similarities across different urban contexts. This vignette applies clustering to identify urban morphological forms and characterises cities by their mix of types.

8.2. SOAR Metrics Utilised

Morphology features (8 variables at 200m scale, aggregated to nearest adjacent street segments):

- **Density**: Building count, Block count
- **Verticality**: Mean building height (median), Height variation (MAD)
- **Scale**: Building footprint area (median)
- **Form complexity**: Fractal dimension (median)
- **Aggregation**: Block coverage ratio (median), Shared walls ratio (median)

External characterisation variables:

- Population density (persons/km^2)
- Street network density (street segment count at 1,200m)
- Land-use diversity (Hill number $q = 0$ at 200m)

8.3. Methodology

We develop a node-level clustering approach that identifies morphological neighbourhood types across all European cities, then profiles each city and country by their distribution across these types.

8.3.1. Clustering Workflow

The workflow (`paper_research/code/eg6_typology/`) operates at the street network node level:

1. **Data preparation**: Extract 8 morphology features at 200m scale for all nodes. Apply log transformation to normalise skewed distributions, then standardise (z-score) across all nodes.

2. **BIRCH clustering**: Apply Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) to identify $k = 8$ morphological neighbourhood types. BIRCH provides O(n) complexity suitable for millions of nodes, building a CF-Tree structure with final agglomerative clustering (Ward linkage) on subclusters.

3. **External characterisation**: For each cluster, compute mean population density, street network density, and land-use diversity to interpret what each morphological type represents in functional terms.

4. **City/country profiling**: Compute the proportion of nodes in each cluster for every city and country, creating compositional vectors that characterise urban form distributions.

5. **Contrasting cluster visualisation**: Identify clusters with highest/lowest density and mixed-use values; plot cities by their proportions in these contrasting types to reveal national patterns.

8.3.2. Feature Selection Rationale

The 8 features capture complementary dimensions of urban form: density metrics (building and block counts) describe how much is built; verticality metrics (height median and variation) describe the skyline profile; scale (building area) captures footprint size; form complexity (fractal dimension) distinguishes regular from irregular building shapes; and aggregation metrics (block coverage, shared walls) describe how buildings relate to their parcels and neighbours. This parsimonious set avoids redundancy while covering the key morphological dimensions identified in urban morphometrics literature.

8.4. Results

Analysis of street network nodes across European cities identifies 8 distinct morphological neighbourhood types with interpretable characteristics.

The resulting cluster profiles suggest distinct urban fabrics. Figure 12 shows standardised feature profiles for each cluster, and Table 11 summarises their key characteristics. Cluster 7 represents dense, tall, complex urban cores with high building counts and shared walls. Cluster 1 represents sparse, low-rise peripheral development with large building footprints but low coverage. Intermediate clusters capture suburban forms, industrial areas, and historic centres with varying combinations of density, height, and complexity.

Table 11: Characteristics of the 8 morphological clusters. Pop Density in persons/km^2, Network Density in nodes/km^2.

Cluster	Nodes	% Total	Pop Density	Network Density	Mixed Use
1	10,641	5.3	5405	909.9	2.50
2	21,660	10.8	13032	1343.2	6.87
3	16,778	8.4	12662	1277.8	5.08
4	26,133	13.1	6724	1076.9	4.15
5	5,513	2.8	8045	920.8	2.89
6	39,268	19.6	4308	969.3	2.72
7	72,122	36.1	4501	816.0	2.27
8	7,885	3.9	17180	1099.1	4.96

External metrics appear to support cluster interpretations. Ranking clusters by population density, street network density, and land-use diversity (Figure 13) confirms that morphological clustering captures functional urban differences. High-density clusters (Cluster 7) exhibit 5–10× higher population density than low-density clusters (Cluster 1), with corresponding differences in street network connectivity and mixed-use intensity.

National patterns appear to emerge in city compositions. When cities are plotted by their proportions in contrasting cluster types (Figure 14), clear geographic patterns appear. Italian cities cluster toward high proportions in dense, mixed-use morphological types. Dutch and German cities show more

Figure 12: **Morphological cluster profiles.** Radar plots showing standardised feature values for each of 8 clusters. Axes represent: Building Count, Block Count, Mean Height, Height Variation, Building Area, Fractal Dimension, Block Coverage, and Shared Walls. Values are z-scores (0 = mean across all clusters).

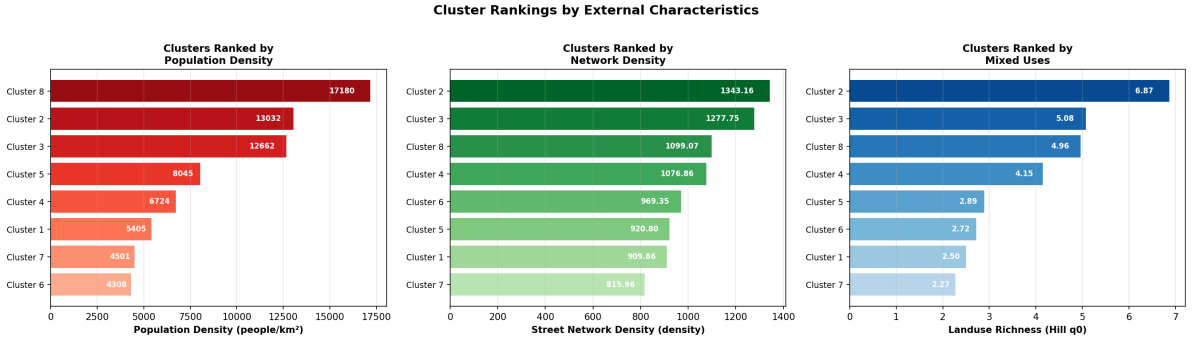

Figure 13: **Clusters ranked by external characteristics.** Horizontal bar charts ranking the 8 morphological clusters by mean population density (left), street network density (centre), and land-use diversity (right). Cluster 7 consistently ranks highest across all three metrics.

balanced distributions. Eastern European cities (Romania, Poland) tend toward higher proportions in lower-density suburban types, potentially reflecting post-socialist development patterns.

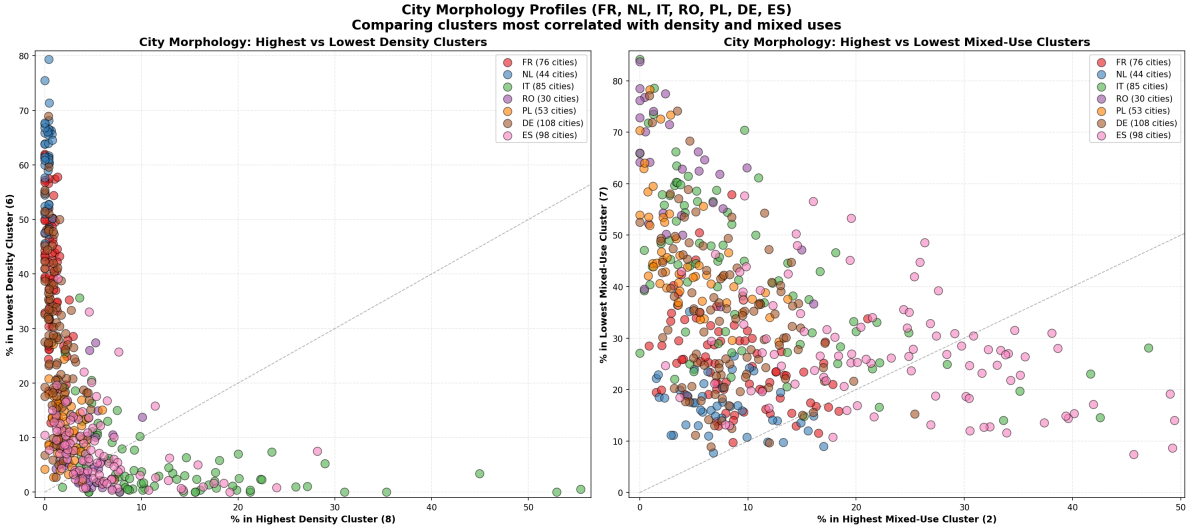

Figure 14: **City morphology profiles by contrasting cluster proportions.** Left: Cities plotted by percentage of nodes in highest-density vs. lowest-density clusters. Right: Cities plotted by percentage in highest vs. lowest mixed-use clusters. Points coloured by country. Italian cities (green) cluster toward dense/mixed forms; Eastern European cities spread toward lower-density types.

Countries cluster by morphological similarity. Hierarchical clustering of country composition vectors (Figure 15) reveals regional groupings: Western European countries (Netherlands, Belgium, Germany) share similar morphological profiles distinct from Southern European (Italy, Spain) and Eastern European (Romania, Poland, Bulgaria) groupings. This suggests that planning traditions, historical development patterns, and regulatory frameworks leave detectable signatures in aggregate urban form.

8.5. Discussion

The analysis suggests that European cities may share common neighbourhood building blocks despite diverse planning traditions. The 8 clusters span a spectrum from sparse peripheral development (Cluster 1: lowest density at 5,405 persons/km^2) to dense urban cores (Cluster 8: highest density at 17,180 persons/km^2). Notably, the most common cluster type (Cluster 7, comprising 36% of sampled nodes) represents moderate-density development with relatively low mixed-use scores, suggesting this suburban-to-urban transitional form dominates European urban fabric.

The external characterisation reveals that morphological clustering captures meaningful functional differences: population density varies by a factor of 4 across clusters, network density by 1.6, and land-use diversity (Hill $q = 0$) by 3. The clusters represent recognisable urban fabrics that recur across national contexts, but the mix varies systematically by country—Greek cities show high concentrations in Cluster 2 (37%), while Dutch and Irish cities concentrate in Cluster 6 (60% and 55% respectively).

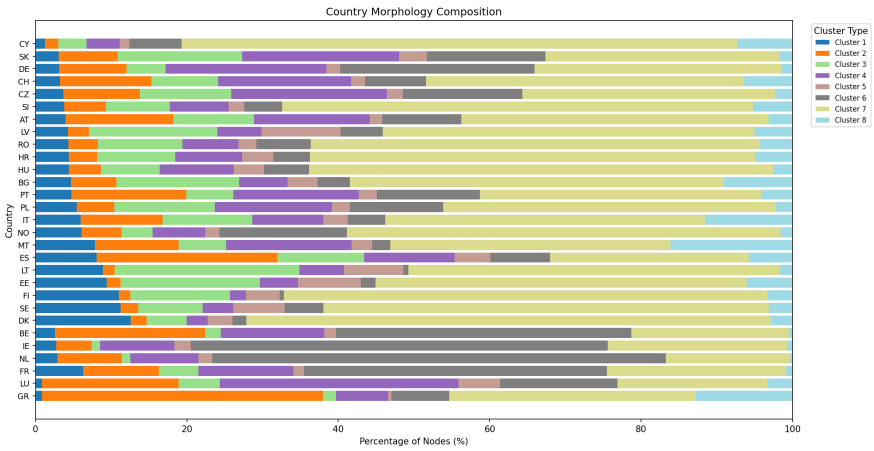

Figure 15: **Country morphology compositions.** Stacked bar chart showing the proportion of nodes in each morphological cluster by country. Countries are ordered by hierarchical clustering of their composition vectors, grouping nations with similar urban form distributions.

Most cities contain multiple morphological types—this heterogeneity would be invisible in city-level aggregates but is captured by the compositional approach. Country clustering suggests that shared planning histories, housing policies, and development economics leave detectable signatures in aggregate urban form. The distinction between Nordic countries (high Cluster 7 proportions), Mediterranean countries (more varied distributions), and Western European countries (Cluster 6 dominant) aligns with known differences in planning traditions.

8.6. Extensions

Potential directions: temporal evolution of compositions using historical imagery; correlations with urban outcomes (energy use, health, economic activity); transitional neighbourhood identification at cluster boundaries; morphological vs functional similarity comparisons; satellite imagery validation of cluster assignments.

8.7. Reproducibility

Code, outputs, and documentation: `https://github.com/[repo]/paper_research/code/eg6_typology/`

9. Vignette 7: Site Selection

9.1. Motivation

Large-scale datasets can be filtered by multiple criteria to identify candidate locations for new facilities, housing, or infrastructure investments. This vignette identifies locations with high centrality, mixed uses, and transport access, but lower population density as potential candidates for development.

9.2. SOAR Metrics Utilised

- **Diversity indices**: Hill numbers ($q = 0$, $q = 1$, $q = 2$) at 400m measuring land-use mix
- **Network centrality**: `cc_beta_1600` (20-minute catchment)
- **Transport access**: `cc_transportation_nearest_max_1600` (distance to nearest stop)
- **Population density**: `density` (persons/km²)
- **Saturation classification**: From Vignette 1 (cities with `Consistently Saturated` or `Variable Saturated` POI coverage)

9.3. Methodology

We classify street network nodes into typologies based on combinations of centrality, diversity, transport access, and density. Nodes with high centrality, high diversity, and good transport access but lower population density represent potential development opportunities—locations where urban infrastructure supports intensification but current utilisation is low.

9.3.1. Classification Workflow

The workflow (`paper_research/code/eg7_site_selection/`) operates at the street network node level:

1. **City filtering**: Select cities with reliable POI coverage (as identified in Vignette 1).
2. **Metric extraction**: For each node, extract diversity indices, centrality, transport distance, and population density.
3. **Threshold classification**: Define high/low thresholds using within-city percentiles (70th percentile for high, 30th for low). Classify nodes as:
   - Mixed-use dense: High diversity and high density
   - Mixed-use opportunity: High diversity, high centrality, good transport access, but low density
   - Single-use dense: Low diversity but high density
   - Peripheral: Low centrality and low diversity
4. **City profiling**: Compute the percentage of nodes in each typology per city.
5. **Ranking**: Rank cities by proportion of mixed-use nodes and by proportion of opportunity nodes.

9.4. Results

Analysis of cities with reliable POI coverage suggests variation in urban form typologies and development potential.

Mixed-use cities have high proportions of nodes with diverse land-use accessibility, indicating fine-grained mixing of residential, commercial, and service functions. Table 12 shows cities with highest mixed-use proportions.

Table 12: Cities with highest proportion of mixed-use nodes.

City	Country	Mixed Score	% Mixed Dense	% Mixed Opp.
Den Haag	NL	0.42	30.0	0.0
Leiden	NL	0.42	30.0	0.0
Venezia	IT	0.41	8.3	3.4
La Rochelle	FR	0.40	30.0	0.0
Grasse	FR	0.40	30.0	0.0
Bayonne	FR	0.39	30.0	0.0
Rætebøl	DK	0.38	30.1	0.0
Rimini	IT	0.38	16.6	1.0
Alessandria	IT	0.37	21.7	0.2
Modena	IT	0.37	18.5	1.2

Opportunity cities have substantial proportions of nodes that combine high connectivity and diversity with lower current population density, suggesting potential for sustainable densification. Table 13 highlights cities with highest opportunity proportions.

Table 13: Cities with highest proportion of development opportunity nodes.

City	Country	% Opportunity	Centrality	Transport
Grasse	FR	21.3	0.37	0.50
Draguignan	FR	19.8	0.39	0.50
Grasse	FR	19.3	0.37	0.50
Bayonne	FR	18.7	0.49	0.50
Cherbourg	FR	18.5	0.50	0.50
Vannes	FR	17.6	0.56	0.50
Søholt	DK	17.1	0.48	0.50
Rætebøl	DK	17.0	0.48	0.50
Toulon	FR	16.9	0.44	0.50
La Rochelle	FR	16.3	0.51	0.50

9.5. Discussion

The typology classification illustrates how to identify locations where infrastructure (connectivity, land-use mix, transport access) exceeds current utilisation (population density). These are not recommendations for development—such decisions require local planning knowledge, market analysis, and community input—but rather a filtering mechanism to identify areas warranting further investigation.

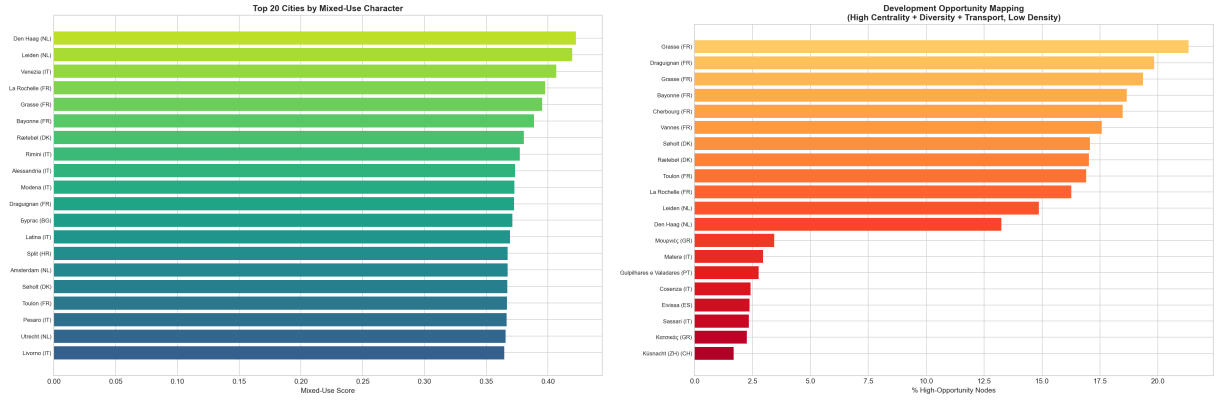

Figure 16: **City rankings by urban form typology.** Left: Cities ranked by proportion of mixed-use nodes. Right: Cities ranked by proportion of development opportunity nodes (high connectivity and diversity with lower population density).

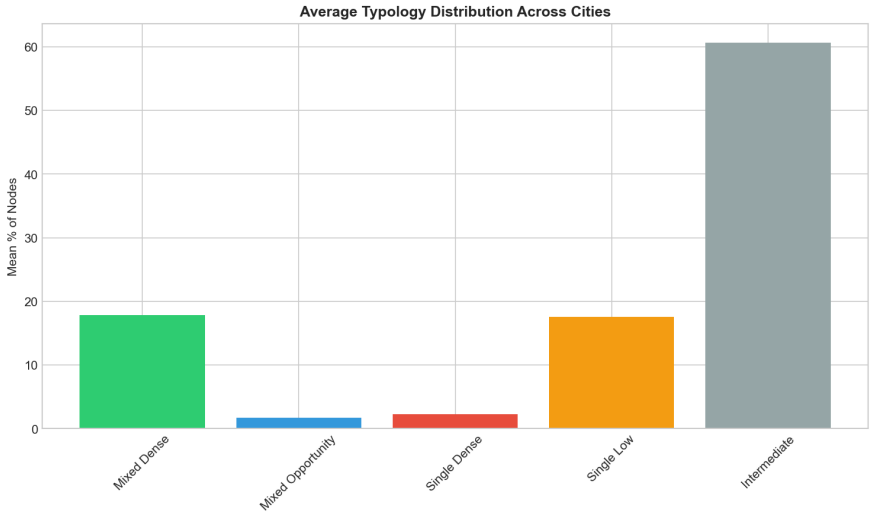

Figure 17: **Distribution of urban form typologies across analysed cities.** Histogram showing the proportion of nodes classified into each typology category (mixed-use dense, mixed-use opportunity, single-use dense, peripheral).

The threshold-based classification is sensitive to parameter choices; different percentile cutoffs would identify different opportunity areas.

9.6. Extensions

Potential directions: additional criteria (land ownership, zoning, building age); scenario modelling for new infrastructure; correlation with property values or vacancy rates; longitudinal analysis of densification patterns; case studies of identified opportunity areas.

9.7. Reproducibility

Code, outputs, and documentation: `https://github.com/[repo]/paper_research/code/eg7_site_selection/`

10. Discussion

10.1. Cross-Cutting Themes

Across all seven vignettes, several themes emerge: (1) data quality assessment provides a foundation for comparative analysis; (2) multi-scale metrics capture neighbourhood effects at varying radii; (3) node-level granularity identifies within-city inequities missed by coarse zonal aggregations; and (4) reproducible workflows using standardised metrics enable researchers new to spatial network analysis.

10.2. Limitations

Each question provides sufficient methodological detail to enable replication. Researchers can expand these analyses with:

- **Domain-specific theoretical frameworks**: Grounding analyses in urban planning theory, geography, sociology, economics, or other relevant disciplines
- **Additional validation**: Incorporating field observations, administrative data, surveys, or behavioral data to test whether patterns hold beyond the available metrics
- **Sensitivity analyses**: Examining how results change with different parameter choices, spatial scales, or methodological approaches
- **Longitudinal perspectives**: Adding temporal dimensions to understand how patterns evolve
- **Contextual depth**: Conducting detailed case studies of specific cities or regions to understand local mechanisms
- **Cross-dataset integration**: Combining SOAR with other data sources (mobility data, economic indicators, policy records) for richer analyses

Additional limitations include: (1) POI data quality variations across regions (addressed in Vignette 1); (2) temporal constraints (SOAR represents a snapshot); (3) lack of behavioural validation (network distances are proxies for actual travel behaviour); (4) computational requirements; and (5) the inherent limitations of any single dataset in capturing urban complexity.

10.3. Adapting These Analyses

Researchers can adapt these analyses by:

- **Parameter tuning**: The spatial scales, distance thresholds, and statistical cutoffs used here are starting points; sensitivity testing may reveal more appropriate values for specific contexts
- **Local data integration**: Combining SOAR with municipal datasets, regional surveys, or national statistics can provide validation and additional explanatory power
- **Methodological alternatives**: The analytical approaches demonstrated here (Random Forests, correlations, descriptive statistics) are illustrative; researchers should explore alternative methods (hierarchical models, spatial econometrics, machine learning ensembles) as appropriate
- **Geographic focus**: While we analyze 699 cities, in-depth investigations of subsets (single countries, specific typologies, matched pairs) may yield richer insights
- **Stakeholder engagement**: Collaborating with planners, policymakers, or community organizations can ensure that analyses address real-world priorities and benefit from local knowledge
- **Computational considerations**: Some analyses may benefit from high-performance computing resources, spatial databases, or cloud platforms

11. Conclusion

This paper presents exploratory worked examples using the SOAR urban data model. These vignettes illustrate the types of analysis the dataset can support—data quality filtering, multi-scale analysis, access gap identification, predictive modelling, benchmarking, typology classification, and site selection—without claiming definitive findings on any urban phenomenon.

The vignettes are starting points. Researchers interested in rigorous investigation of the questions raised may need to:

- Ground analyses in domain-specific theory
- Validate patterns against local data sources
- Conduct sensitivity analyses on parameters and thresholds
- Consider causal mechanisms rather than correlational patterns

The contribution is practical: reproducible code and clear workflows that lower the barrier to entry for researchers evaluating whether SOAR suits their needs.

Acknowledgements

This work was supported by the European Union's Horizon programme under the TWIN2EXPAND project. We acknowledge the data providers: Overture Maps Foundation, Copernicus Land Monitoring Service, and Eurostat.

Funding

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

References

[1] G. Simons, Others, Soar: A scalable, open, automated, and reproducible urban data model for the eu, Data in BriefIn preparation (2025).

[2] G. Boeing, Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks, Computers, Environment and Urban Systems 65 (2017) 126–139.

[3] 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).
URL https://discovery.ucl.ac.uk/id/eprint/10134012/

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

[5] M. Fleischmann, momepy: Urban morphology measuring toolkit, Journal of Open Source Software 4 (43) (2019) 1807.

[6] G. Simons, The cityseer python package for pedestrian-scale network-based urban analysis, Environment and Planning B: Urban Analytics and City Science 49 (9) (2022) 2356–2361. doi:10.1177/23998083221133827.