Ten Exploratory Questions 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 examine POI data quality, green space accessibility, educational infrastructure, amenity prediction, 15-minute city benchmarking, and urban morphology clustering. Each vignette provides a reproducible workflow and suggests extensions. The purpose 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 [TODO:], 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 (400–9,600m): 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 urban phenomenon. The vignettes are intentionally simple, using standard methods (correlations, Random

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

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 vignettes are sequenced by analytical complexity. Vignette 1 assesses POI data quality and identifies cities with reliable coverage. Subsequent vignettes address equity (green space access), infrastructure gaps (education, transit), benchmarking (15-minute cities), predictive modelling (POI demand, densification potential), and comparative geography (cross-national patterns, urban typologies).

The remainder of this paper is structured as follows: Section 2 reviews related work on urban data applications; Sections 3–12 present the ten vignettes; Section 13 discusses cross-cutting themes and limitations; Section 14 concludes.

2. Related Work

[TODO: Brief review of: (1) multi-scale urban datasets (OSMnx, Urban Observatory, etc.); (2) POI quality assessment methods; (3) comparative urban analysis frameworks; (4) walkability and accessibility metrics; (5) urban typology clustering approaches. 2-3 pages.]

3. Vignette 1: How Can We Assess POI Data Quality Across Cities?

3.1. Motivation

Point of interest (POI) datasets derived from crowdsourced platforms like OpenStreetMap exhibit spatially heterogeneous completeness, with systematic underrepresentation in peripheral regions and developing economies. Comparative analyses using raw POI counts risk conflating true urban form differences with data quality artefacts. Before conducting cross-city comparisons, researchers must identify which cities have sufficiently complete POI coverage to support reliable analysis.

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/poi_saturation_notebook.py`) 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$^{\beta}$), 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

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.

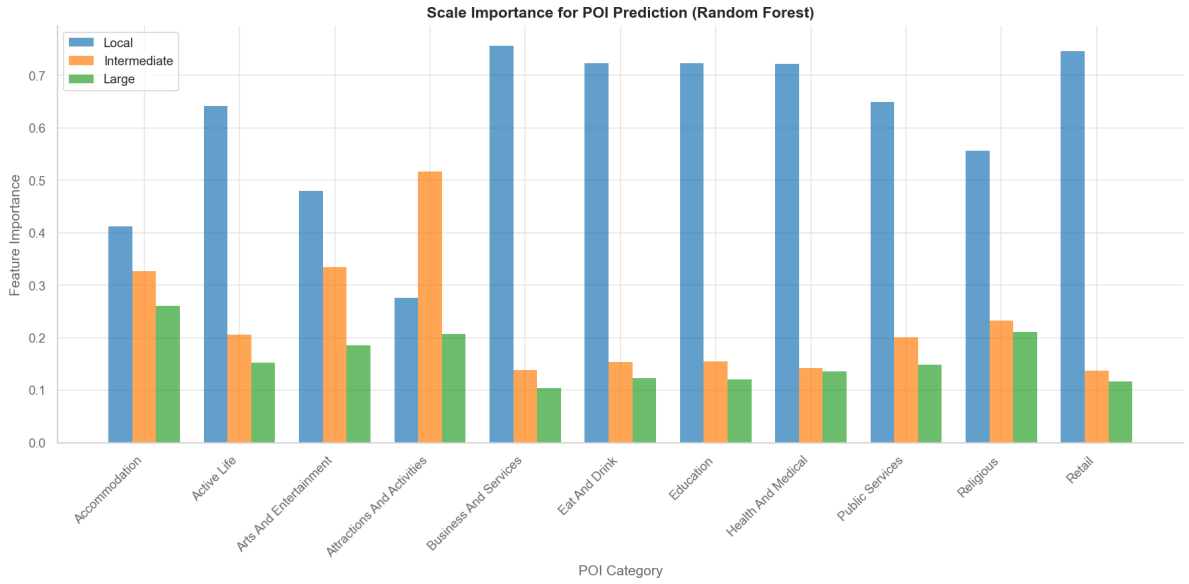

Figure 1: Feature importance showing which population scale (local, intermediate, large) best predicts POI distribution for each category.

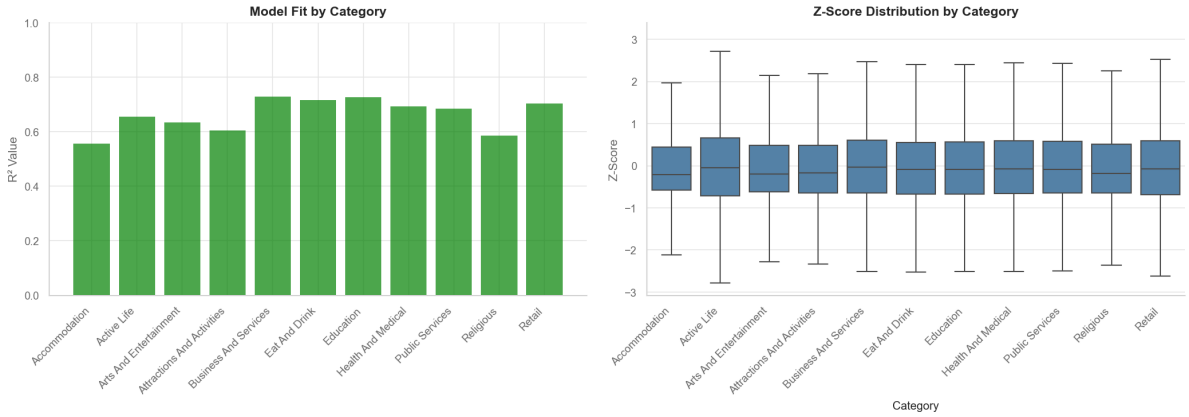

Figure 2: Exploratory data analysis. Left: Random Forest model fit (R^2) by POI category. Right: distribution of z-scores across grid cells per category.

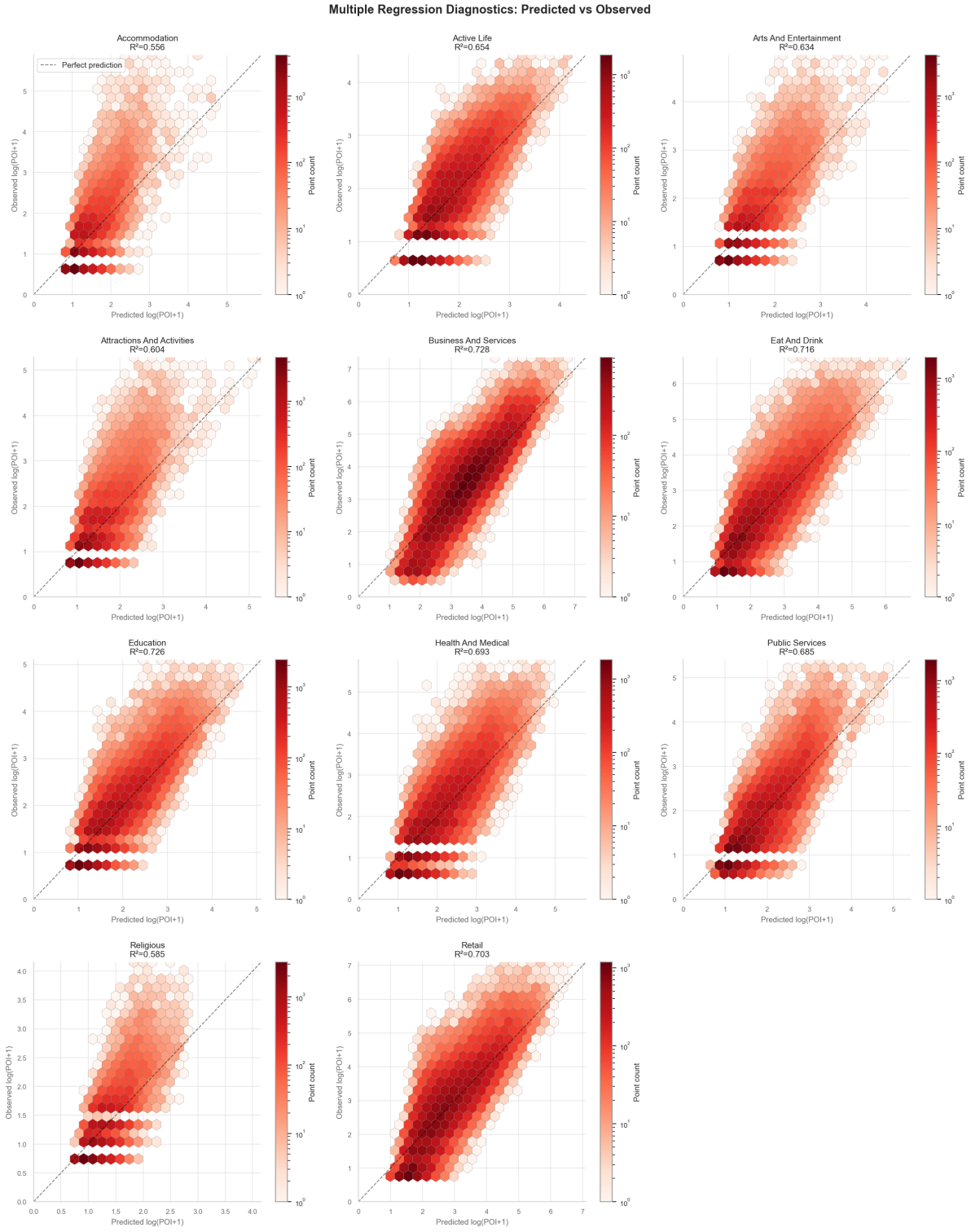

Figure 3: Regression diagnostics: predicted vs. observed POI counts (log scale) for each category.

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.

3.6. 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_poi_saturation/`

4. Vignette 2: Does Urban Density Compromise Green Space Access?

4.1. Motivation

Do denser neighbourhoods have better or worse access to parks and tree canopy? This vignette examines within-city correlations between population density and green space proximity across European cities.

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 491 cities across 18.7 million street network nodes reveals a consistent within-city pattern for green blocks alongside contrasting behavior for tree canopy:

Green space (parks): 487 cities (99%) exhibit positive correlations, where denser areas face longer walks to parks. Median distance is 70.7m, with 91.1% of nodes within a 5-minute walk (400m). The strongest positive correlation (Verviers, Belgium: $\rho = 0.76$) exemplifies peripheral park placement, while rare negative outliers like Meiderich/Beeck, Germany ($\rho = -0.06$) and Spijkenisse, Netherlands ($\rho = -0.04$) demonstrate integrated green infrastructure in high-density zones.

Tree canopy: 478 cities (97%) show negative correlations, indicating that denser neighbourhoods have *better* tree canopy access. Median distance is 76.6m, with 85.9% within 400m. Strong negative correlations (e.g., Soest, Netherlands: $\rho = -0.69$; Rüsselsheim am Main, Germany: $\rho = -0.62$) suggest street tree programmes concentrated in urban cores, likely reflecting municipal maintenance priorities and sidewalk infrastructure availability.

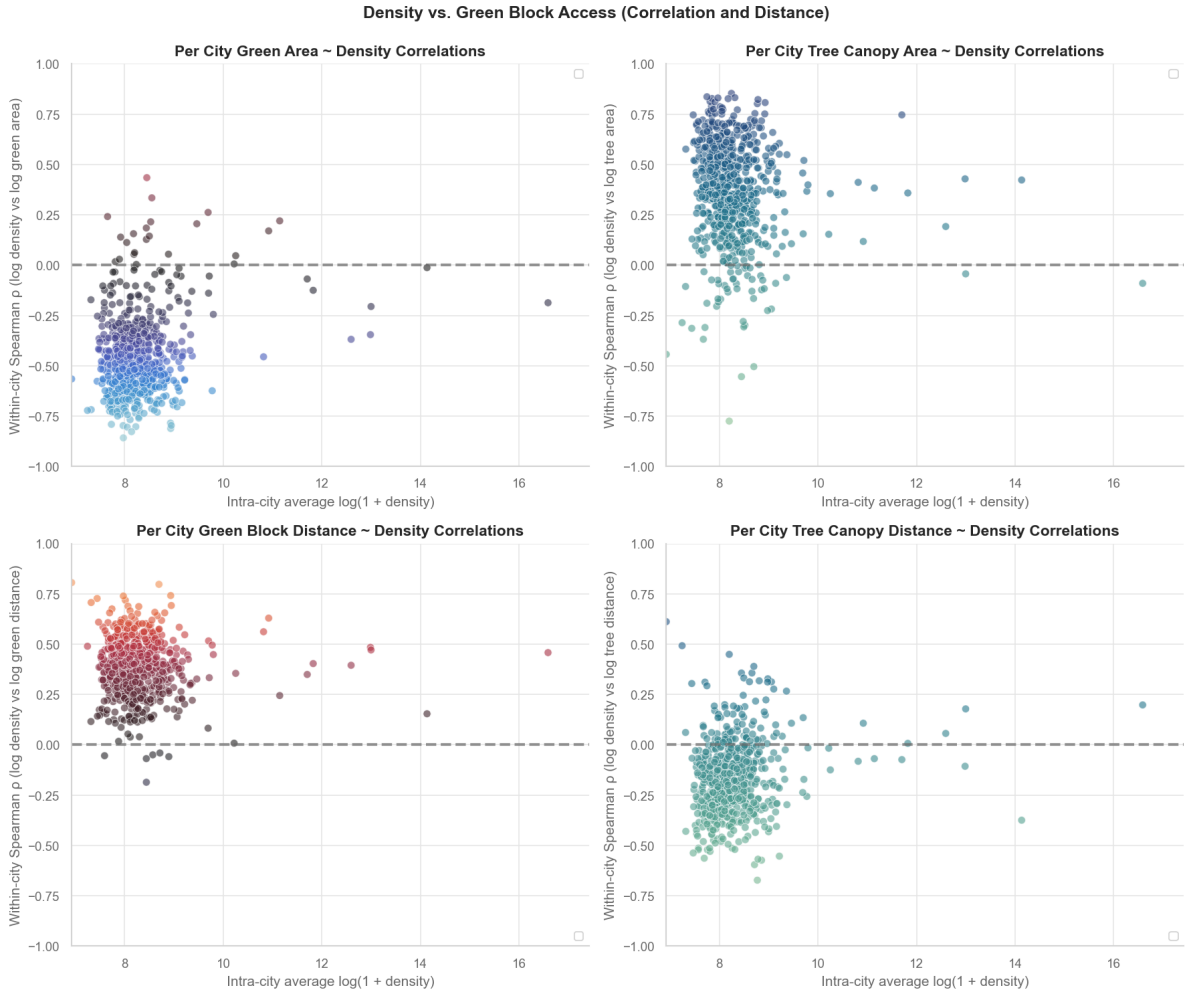

Figure 5: **Green space accessibility and tree canopy versus population density.** 2×2 grid comparing distance metrics (top row) and correlation analysis (bottom row) across 491 European cities. Left column: green blocks (parks). Right column: tree canopy. Points colored by Spearman correlation strength (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.

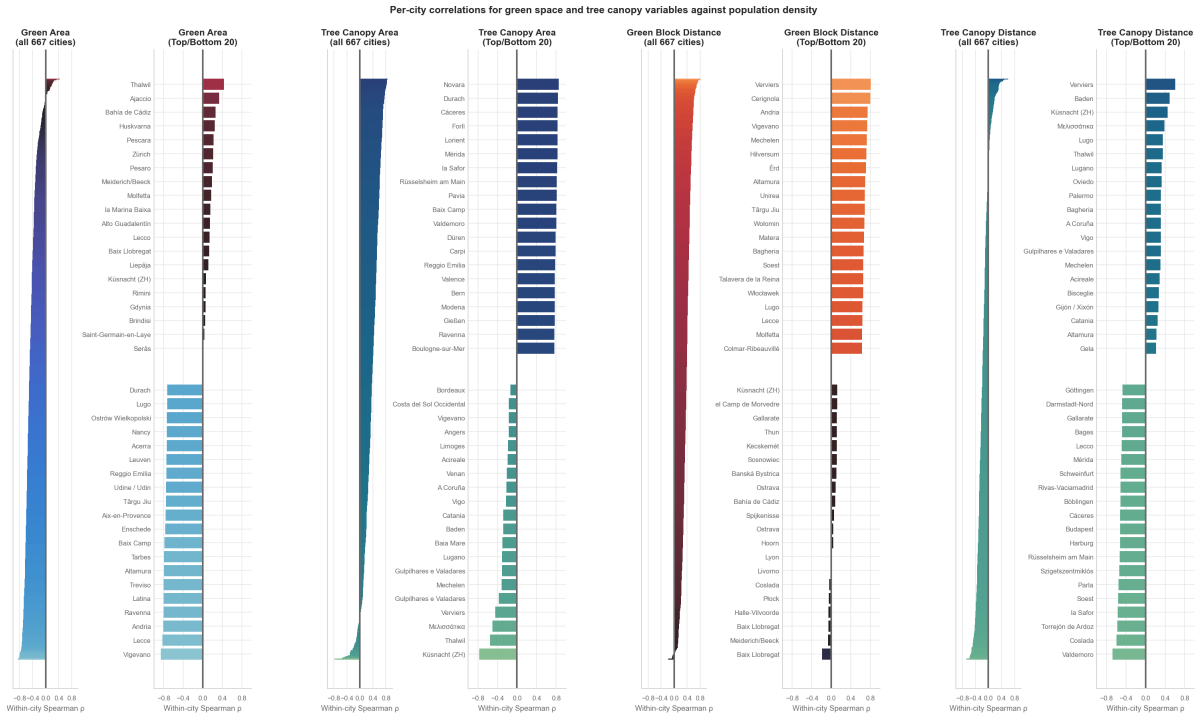

Figure 6: **Per-city correlation patterns for green space accessibility.** Diverging bar chart ranking cities by density-green correlation. For green blocks: only 4 cities show negative correlations (dense neighborhoods closer to parks); 487 cities show positive correlations (parks in peripheries). For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

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_green_space/`

5. Vignette 3: Where Are Educational Infrastructure Gaps Most Pronounced?

5.1. Motivation

Access to educational facilities is a fundamental urban equity issue, directly affecting the daily lives of families and children. The spatial distribution of educational facilities varies widely across European cities. We restrict this analysis to cities with `Consistently Saturated` education POI coverage (see Question 1).

5.2. SOAR Metrics Utilised

- `cc_education_nearest_max_1600`: Network distance to nearest education POI
- `cc_education_1600_wt`: Weighted count of education POIs within 1,600m
- Census-derived population (per-node denominators)

5.3. Methodology

For each city, 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 (comparing the 75th and 25th percentiles) and the percentage of nodes with access worse than twice the city mean. Analysis is restricted to cities with stable POI coverage (as identified in Question 1), ensuring that results reflect genuine service gaps rather than data artefacts.

Table 2: Best and worst access to education (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

5.4. Results

Access to education is a tale of two Europes. In cities like Küsnacht (CH) and Płock (PL), over 65% of nodes are within a 5-minute walk of a school, and mean access distances are under 400m. By contrast, in cities like Como (IT) and Iserlohn (DE), mean distances exceed 600m and fewer than 40% of nodes are within 400m. Table 2 highlights the top and bottom performers.

Equity is not guaranteed by abundance. Even in cities with good average access, pockets of disadvantage persist. The P75/P25 ratio ranges from 2.4 (Nieuwegein, NL) to over 6 (Xánthi, GR), and in the least equitable cities, nearly one in five nodes is severely underserved (Table 3).

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

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 dramatically across Europe. The P75/P25 ratio distinguishes cities with equitable distribution from those with concentrated provision. High-performing cities (Venezia, Almere) combine compact form with distributed school placement. Underperforming cities often exhibit suburban sprawl or consolidated school networks. Equity metrics reveal that average access can mask substantial within-city disparities.

5.6. Extensions

Potential directions: school capacity and enrolment data; socioeconomic correlates; temporal trend analysis; case studies of high-performing cities; school location scenario modelling.

5.7. Reproducibility

Code, outputs, and documentation: `https://github.com/[repo]/paper_research/code/eg3_education/`

6. Vignette 4: Can We Predict Amenities using Network Centrality and Census Data?

6.1. Motivation

Can network centrality and census demographics predict where amenities locate? This vignette trains regression models to predict POI counts from structural features, testing how well commercial location correlates with network accessibility across different cities.

6.2. SOAR Metrics Utilised

- **Network centrality**: Closeness and betweenness centrality at 400m, 800m, 1,200m, 1,600m, 4,800m, and 9,600m 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 Question 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 Question 1), yielding 21 cities and 999,180 street network nodes.

6.3.1. Model Training Workflow

The workflow (`paper_research/code/eg4_amenity_prediction/`) operates at the street network node level:

1. **Data preparation**: Extract network centrality metrics at six spatial scales (400–9,600m) 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 12 centrality features (closeness and betweenness at 6 scales) 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, minimum samples per leaf of 50.

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: $R^2 = 0.723$ for eat & drink, $R^2 = 0.724$ for business & services. Per-city R^2 values reveal how consistently the centrality-amenity relationship holds across different urban forms.

Eat & drink: Median city $R^2 = 0.724$, with 90.5% of cities achieving $R^2 > 0.5$. Italian cities dominate the best-predicted list (Table **??**), suggesting that network structure strongly determines hospitality location in these contexts. Gallarate ($R^2 = 0.35$) and Heerlen ($R^2 = 0.45$) show poorest fit, potentially indicating amenity distributions driven by factors beyond network accessibility.

Business & services: Median city $R^2 = 0.693$, with 95.2% of cities exceeding $R^2 > 0.5$. The pattern mirrors eat & drink, with Italian cities showing strongest network-amenity alignment (Table **??**).

City	R^2	MAE	RMSE	Nodes
Bari	0.857	0.430	0.584	14,855
la Safor	0.822	0.423	0.601	5,435
Ragusa	0.819	0.374	0.511	6,693
la Plana Alta	0.817	0.446	0.627	10,265
Alessandria	0.809	0.457	0.590	5,417
...				
Heerlen	0.439	0.488	0.612	33,426
Pordenone / Pordenon	0.582	0.499	0.657	10,823
Gallarate	0.585	0.479	0.606	17,720
Bergamo	0.590	0.496	0.638	41,833

City	R^2	MAE	RMSE	Nodes
Ragusa	0.865	0.406	0.570	6,693
Cremona	0.816	0.499	0.643	6,138
Bari	0.810	0.548	0.724	14,855
la Safor	0.804	0.550	0.759	5,435
Alessandria	0.799	0.535	0.710	5,417
...				
Modena	0.565	0.679	0.892	14,782
Heerlen	0.595	0.566	0.726	33,426
Pordenone / Pordenon	0.623	0.600	0.771	10,823
Prato	0.636	0.639	0.797	20,100

Feature importance reveals that intermediate-scale closeness centrality (1,200–1,600m) dominates predictions for both categories (Table **??**), 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.

Eat & Drink		Business & Services	
Feature	Importance	Feature	Importance
Cc Beta 1200	0.171	Cc Beta 1200	0.214
Cc Beta 1600	0.157	Cc Beta 1600	0.169
Y 1564	0.117	Cc Beta 800	0.147
Cc Beta 800	0.115	Y 1564	0.100
Y Ge65	0.092	Emp	0.074
Density	0.085	Y Lt15	0.068
Emp	0.083	Y Ge65	0.059
Y Lt15	0.069	Density	0.053
Cc Beta 4800	0.053	Cc Beta 4800	0.045
Cc Beta 400	0.031	Cc Beta 400	0.040

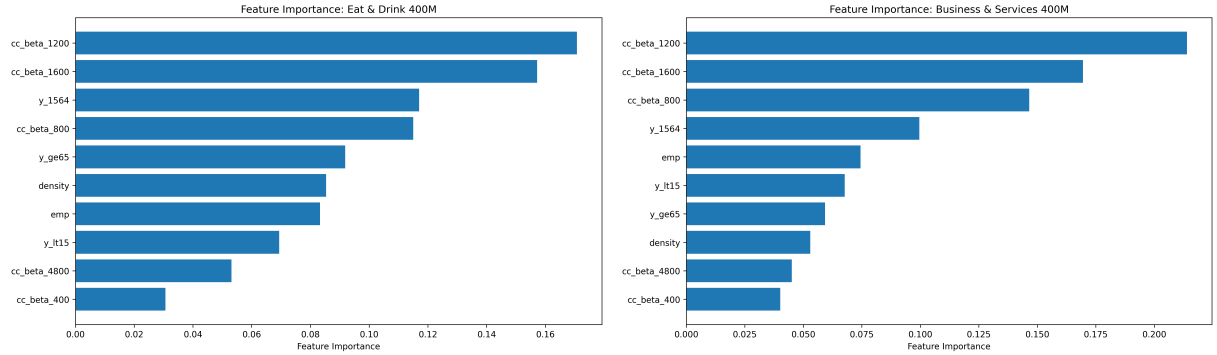

Figure 7: **Feature importance for amenity prediction models.** Left: Eat & drink. Right: Business & services. Intermediate-scale closeness centrality (1,200–1,600m) dominates both models, consistent with pedestrian catchment theory.

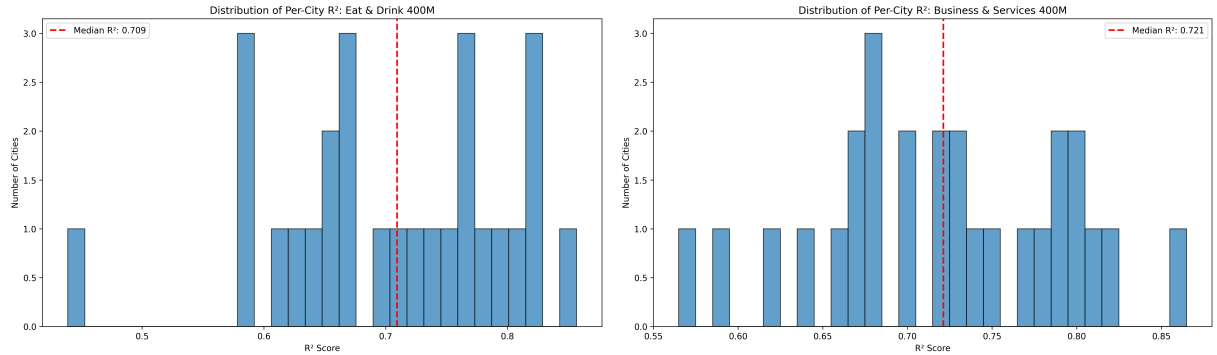

Figure 8: **Distribution of per-city R^2 scores.** Left: Eat & drink. Right: Business & services. Most cities cluster at high R^2 values, with a few outliers showing weaker network-amenity relationships.

6.5. 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.

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_amenity_prediction/`

7. Vignette 5: Which Cities Best Approximate the 15-Minute City Ideal?

7.1. Motivation

The "15-minute city" concept proposes that residents should access essential services within a 15-minute walk. This vignette operationalises the concept using SOAR's pre-computed POI distances, computing what proportion of street network locations have access to all 10 POI categories within 1,200m.

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 Question 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 Question 1's between-category quadrant classification).

7.3.1. Completeness Scoring Workflow

The workflow (`paper_research/code/eg5_poi_minutes/`) 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 (Question 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 reveals substantial variation in 15-minute city completeness across European urban centres.

Few cities achieve true 15-minute completeness. The median city has only a modest percentage of nodes with access to all 10 POI categories within 1,200m. This finding underscores that the 15-minute city ideal remains aspirational for most European cities—even those with good overall accessibility may lack universal coverage across all service types.

Top-performing cities demonstrate that compact urban form and mixed-use development can deliver near-complete 15-minute access. Table **??** 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.

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 exhibit either sprawling urban form, car-oriented development patterns, or systematic gaps in specific service categories. Table **??** highlights cities with lowest completeness scores.

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

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

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

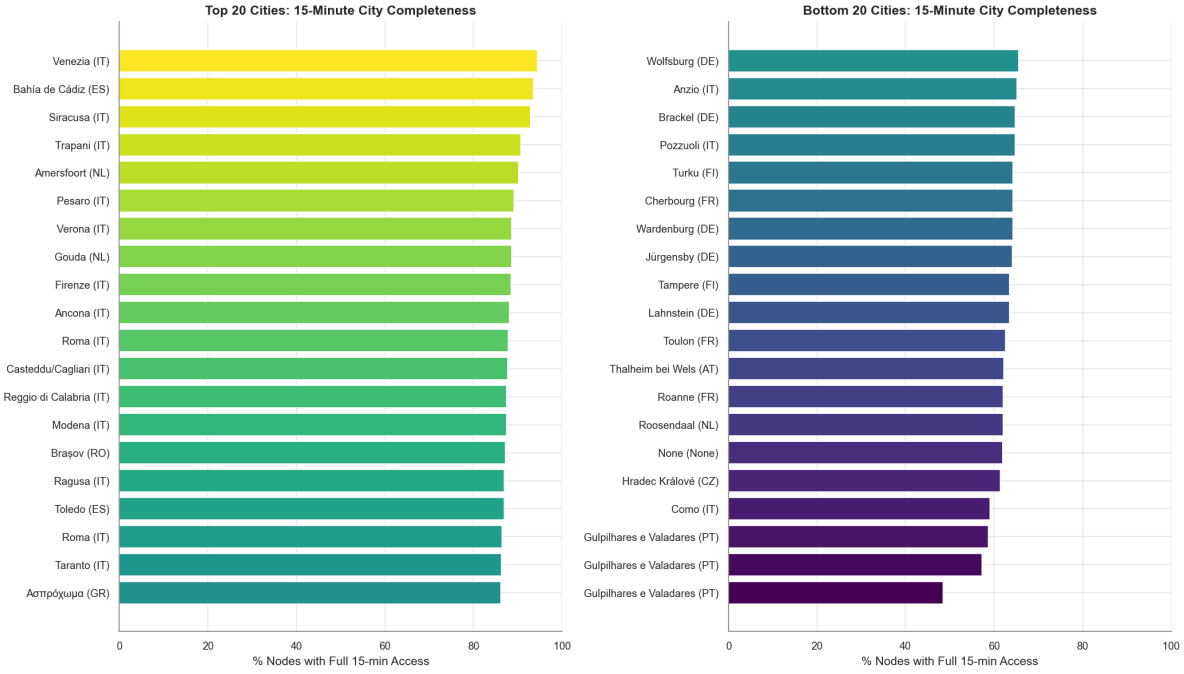

Figure 9: **15-minute city completeness ranking.** Left: Top 20 cities with highest proportion of nodes achieving full 15-minute access. Right: Bottom 20 cities. Colour intensity reflects completeness score.

7.5. Discussion

Few cities have high proportions of nodes with access to all 10 categories. Full access is typically limited by one or two categories rather than general deficits. This metric is sensitive to category definitions and distance thresholds; different operationalisations would yield different rankings.

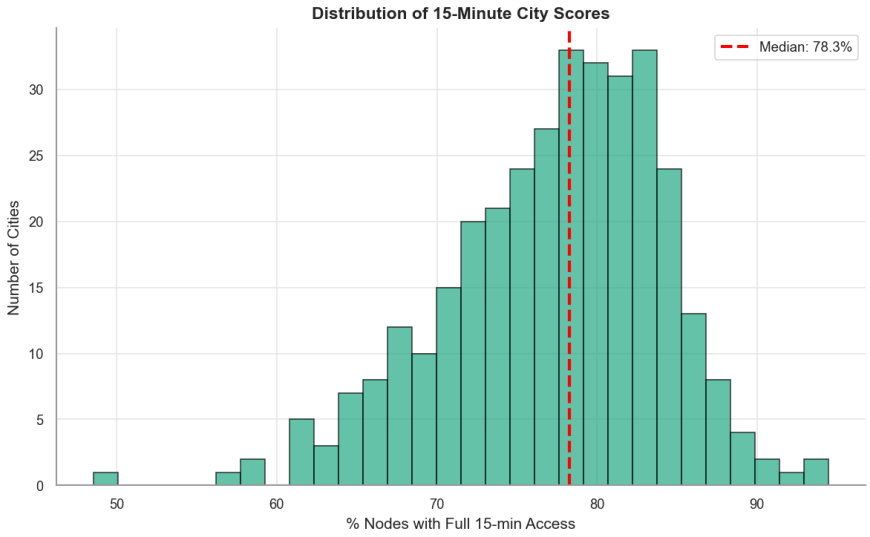

Figure 10: **Distribution of 15-minute city scores across European cities.** Histogram showing the percentage of nodes with full access to all 10 POI categories. Red dashed line indicates median.

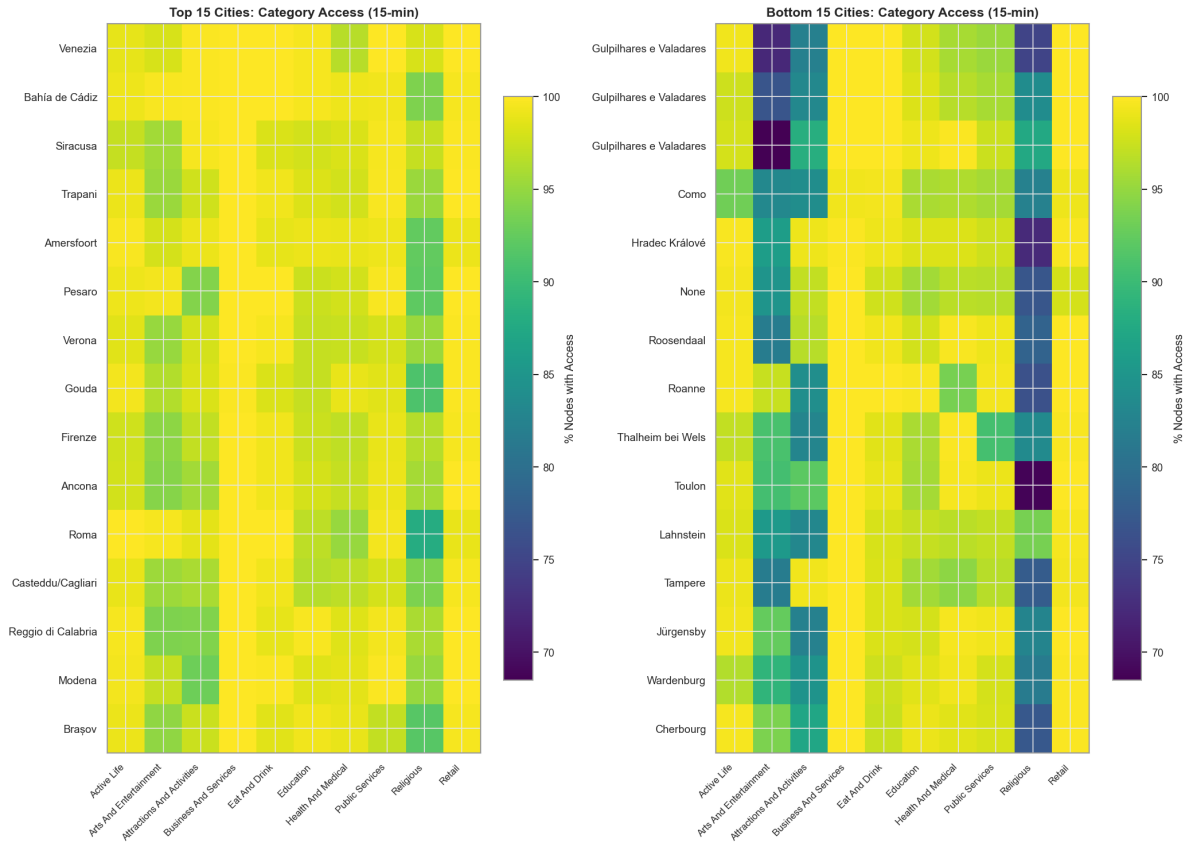

Figure 11: **Per-category access rates for top and bottom cities.** Heatmap showing which POI categories are most/least accessible within 1,200m. Columns represent categories; rows represent cities. Darker colours indicate lower access rates, revealing category-specific bottlenecks.

7.6. Extensions

Potential directions: category weighting by use frequency; actual routing travel times; opening hours analysis; cycling vs walking comparisons; socioeconomic correlates; scenario modelling.

7.7. Reproducibility

Code, outputs, and documentation: `https://github.com/[repo]/paper_research/code/eg5_poi_minutes/`

8. Vignette 6: How Do Morphology Patterns Vary Across European Cities?

8.1. Motivation

Do European cities share common morphological patterns, or does each develop unique forms? This vignette identifies recurring neighbourhood types via clustering, then 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_morphology/`) 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.

Cluster profiles reveal distinct urban fabrics. Figure 12 shows standardised feature profiles for each cluster. 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.

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

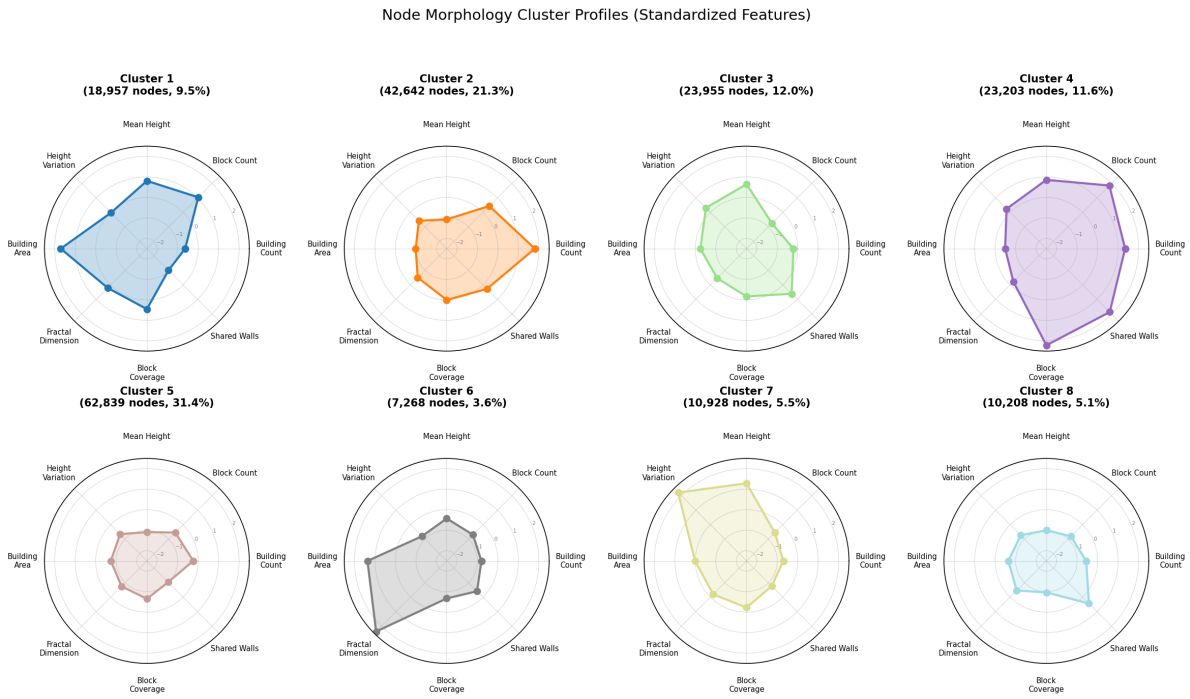

Figure 12: **Morphological cluster profiles.** Radar plots showing standardised feature values for each of 8 clusters. Axes represent: Building Count, Block Count, Mean Height, Height Variation, Building Area, Fractal Dimension, Block Coverage, and Shared Walls. Values are z-scores (0 = mean across all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

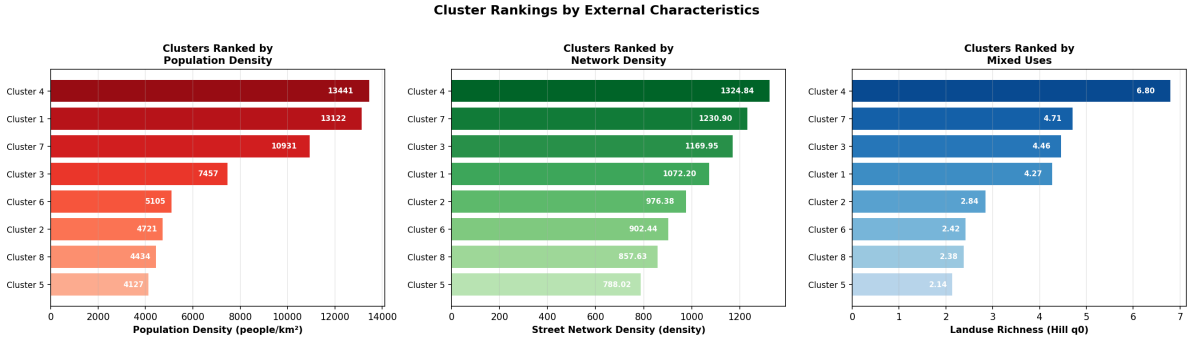

Figure 13: **Clusters ranked by external characteristics.** Horizontal bar charts ranking the 8 morphological clusters by mean population density (left), street network density (centre), and land-use diversity (right). Cluster 7 consistently ranks highest across all three metrics.

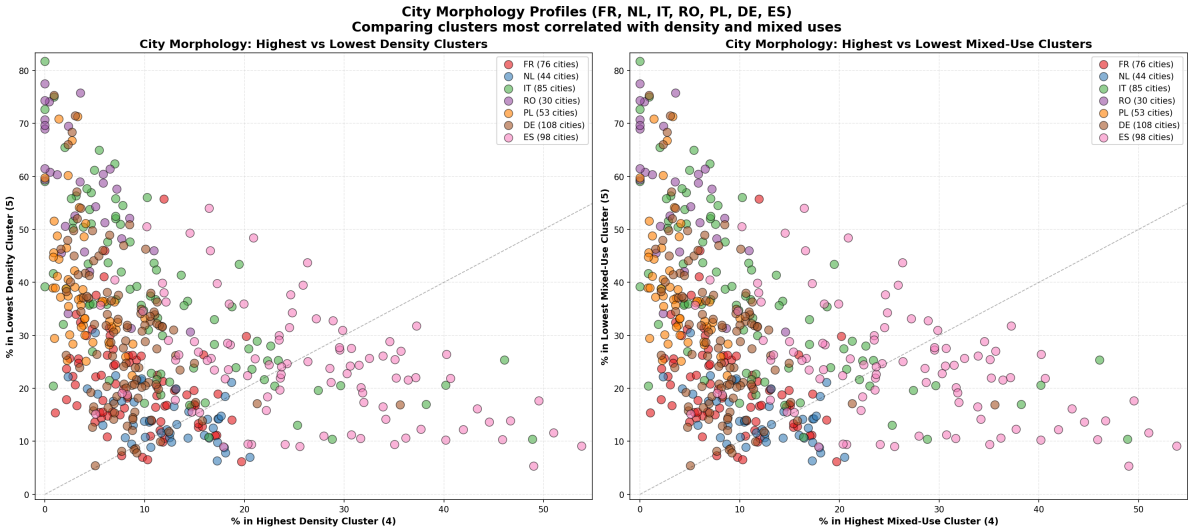

Figure 14: **City morphology profiles by contrasting cluster proportions.** Left: Cities plotted by percentage of nodes in highest-density vs. lowest-density clusters. Right: Cities plotted by percentage in highest vs. lowest mixed-use clusters. Points coloured by country. Italian cities (green) cluster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

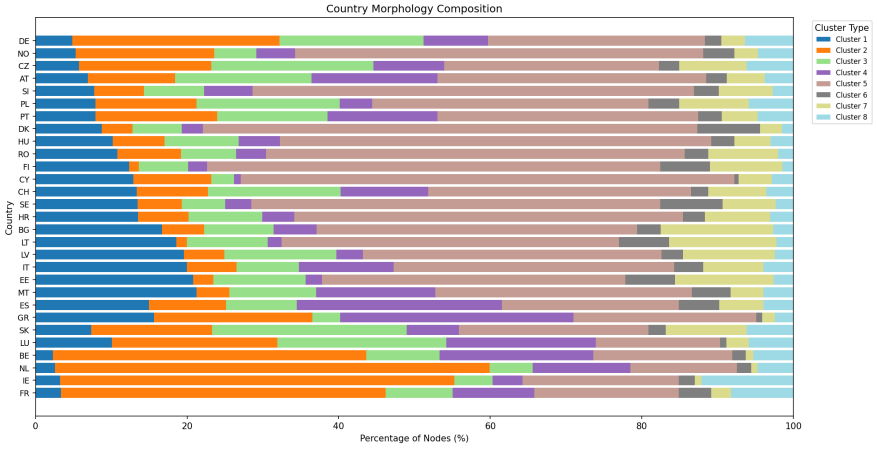

Figure 15: **Country morphology compositions.** Stacked bar chart showing the proportion of nodes in each morphological cluster by country. Countries are ordered by hierarchical clustering of their composition vectors, grouping nations with similar urban form distributions.

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

European cities share common neighbourhood building blocks despite diverse planning traditions. The 8 clusters represent recognisable urban fabrics that recur across national contexts, but the mix varies systematically by country. 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 shared planning histories and development economics create detectable signatures in aggregate urban form.

8.6. Extensions

Potential directions: temporal evolution of compositions; correlations with urban outcomes; transitional neighbourhood identification; morphological vs functional similarity; satellite imagery validation.

8.7. Reproducibility

Code, outputs, and documentation: `https://github.com/[repo]/paper_research/code/eg6_morphology/`

9. Discussion

9.1. Cross-Cutting Themes

Across all eight questions, 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.

9.2. 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 Question 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.

9.3. 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

10. 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 assessment, accessibility comparisons, infrastructure gap identification, predictive modelling, benchmarking, and morphology clustering—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

[TODO: Acknowledge TWIN2EXPAND consortium, funding sources, data providers.]

References