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

Comprehensive urban datasets enable new forms of comparative analysis, yet their potential applications often remain underexplored in data descriptor publications. This paper presents ten exploratory questions demonstrating how the SOAR (Scalable, Open, Automated, and Reproducible) urban data model—a pan-European dataset covering 699 cities with over 100 metrics per street network node—can be applied to diverse urban research domains. We begin with POI data quality assessment using multi-scale Random Forest regression, revealing systematic geographic patterns where Spanish cities show mean z-scores of -0.6 to -1.1 while Central European cities cluster near zero. Subsequent questions examine green space equity, educational infrastructure gaps, 15-minute city benchmarking, and urban typology clustering using SOAR's multi-scale network centrality (400–9,600m), accessibility metrics (400–1,600m), Hill diversity indices, building morphology, and census demographics. Each example provides reproducible workflows and identifies opportunities for extension, offering entry points for researchers to develop more detailed investigations.

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

*Corresponding author

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

1. Introduction

The proliferation of open urban datasets has created opportunities for large-scale comparative analysis. However, large-scale urban datasets can be difficult for researchers to assess from the outside: understanding what data are contained, what questions can be addressed, and how analytical workflows might be structured requires substantial initial investment. To help address this challenge, this paper demonstrates applications of the SOAR (Scalable, Open, Automated, and Reproducible) urban data model [1] through 10 vignettes that illustrate the dataset's contents, the types of questions it can be used to explore, and practical analytical approaches.

The SOAR urban data model provides a standardised framework for 699 European urban centres, combining Eurostat boundaries and demographics, Copernicus Urban Atlas land cover, and Overture Maps infrastructure data. SOAR computes over 100 metrics per street network node at six spatial scales (400–9,600m), encompassing network centrality, land-use accessibility, building morphology, green space proximity, and demographic characteristics.

This paper presents ten analytical vignettes using SOAR to investigate diverse urban phenomena. We adopt the term "vignette" to denote self-contained worked examples that combine motivation, methodology, analysis, and interpretation—each standing alone while contributing to a broader narrative about urban data applications. Importantly, these vignettes are purposely *bite-sized demonstrations* rather than exhaustive treatments. Each vignette showcases what is possible with SOAR and provides a reproducible workflow, but does not attempt encyclopedic depth; extensions suggest directions for more comprehensive analysis. This design enables accessibility for readers new to the dataset while maintaining scientific rigour through transparent methodology and reproducible code. The vignettes cover:

- 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:

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

   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 × 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 pronounced geographic patterns in POI data saturation. **Central and Western European cities consistently perform best**, with German cities (Düren, Iserlohn, Wolfsburg, Jena, Hilden) and Dutch cities (Enschede, Ridderkerk, Veenendaal, Roosendaal) achieving mean z-scores near zero with low spatial variability. French cities (Aix-en-Provence, Perpignan, Tours, Grenoble), Belgian cities (Antwerp), and Italian metropolitan centres (Milano) exhibit similar saturation. These cities demonstrate balanced POI distributions aligning closely with population-based expectations across all 11 land-use categories, indicating reliable data for infrastructure analysis.

Peripheral European regions show systematic undersaturation. Spanish cities dominate the undersaturated category, particularly Madrid satellite municipalities: Parla (−1.10), Valdemoro (−0.81), Alcorcón (−0.81), Fuenlabrada (−0.73), Arganda del Rey (−0.73), Coslada (−0.58), Torrejón de Ardoz (−0.54). Major Spanish cities also underperform (Bilbao −0.61, Basque region −0.68), alongside Spanish exclaves (Ceuta −0.86, Melilla −0.62). Eastern European cities exhibit parallel patterns: Romanian cities (Brăila −0.89, Galați −0.56, Ploiești −0.55, Buzău −0.53), Bulgarian cities (Dobrich −0.87, Sliven −0.64, Haskovo −0.57, Pleven −0.54), Polish cities (Bydgoszcz −0.63), Czech cities (Ostrava −0.58), and Lithuanian cities (Panevėžys −0.53). Italian southern cities (Andria −0.60, Cerignola −0.57) and French peripheral towns (Mantes-la-Jolie −0.58) also show undersaturation.

Nordic countries exhibit **mixed performance**: Swedish cities like Västerås perform well (0.003), while others show deficits. Notably, some saturated cities include: Lithuanian Kaunas (0.001), Slovak Košice (0.003), Croatian Zagreb (0.015), Polish Zielona Góra (0.017), and even Spanish El Bierzo (0.016), indicating heterogeneity within countries.

This **core-periphery pattern** likely reflects: (1) differential OpenStreetMap contributor activity feeding Overture; (2) varying commercial formalisation and business registration practices; (3) regional differences in POI aggregator market coverage. The pattern is pronounced for business services and retail ($\mathbf{R^2}$=0.73, 0.70), where Central European cities show near-complete coverage while Southern/Eastern cities fall 0.5–1.5 standard deviations below expected values. Accommodation shows weakest predictability ($\mathbf{R^2}$=0.56), suggesting tourism infrastructure follows different spatial logic than population-based models predict.

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 & services	0.73	0.76	0.14	0.10
Education	0.73	0.72	0.16	0.12
Eat & drink	0.71	0.72	0.15	0.12
Retail	0.70	0.75	0.14	0.12
Health & medical	0.69	0.72	0.14	0.14
Public services	0.68	0.64	0.21	0.15
Active life	0.66	0.64	0.21	0.15
Arts & entertainment	0.63	0.45	0.36	0.18
Attractions & activities	0.60	0.26	0.53	0.21
Religious	0.59	0.56	0.23	0.21
Accommodation	0.56	0.40	0.34	0.26

Figure 1: Feature importance analysis showing which population scale (local, intermediate, large) best predicts POI distribution for each category. Taller bars indicate stronger predictive power.

Figure 2: Exploratory data analysis. Left: Random Forest model fit (R^2) by POI category, with values ranging from 0.4–0.9 depending on category predictability. Right: distribution of z-scores across grid cells per category, revealing saturation patterns.

Figure 3: Regression diagnostics: predicted vs. observed POI counts (log scale) for each category. Points near the diagonal indicate accurate predictions; outliers represent grid cells with unexpected POI distributions.

Figure 4: City quadrant analysis showing saturation classification. Each of 12 panels represents a POI category (11 categories plus between-category summary). X-axis: mean z-score (saturation level; negative indicates undersaturation). Y-axis: standard deviation of z-scores (spatial variability within city). Quadrant colours: red = consistently undersaturated; green = consistently saturated; orange = variable undersaturated; blue = variable saturated.

3.5. Implications

POI data quality varies systematically across European cities. Researchers comparing walkability or mixed-use metrics between, e.g., German and Spanish cities, may conflate true urban form differences with data artefacts. We recommend: (1) restricting cross-regional analyses to consistently saturated cities; (2) stratifying by saturation quadrant; or (3) applying z-score corrections to accessibility metrics in undersaturated regions.

3.6. Extensions

[TODO: What about using saturation vs category vectors as possible next steps] Future work could explore temporal trends in POI data quality; develop category-specific quality metrics; integrate municipal records or commercial databases for validation; investigate correlations between data quality and urban characteristics (GDP, digital infrastructure, civic engagement); or develop automated quality flagging systems.

3.7. Reproducibility

The analysis generates `grid_counts_regress.gpkg` (grid-level z-scores and predictions), `city_analysis_results.gpkg` (city-level statistics and quadrant classifications), and diagnostic visualizations in `paper_research/code/eg1_poi_compare/outputs/`. Subsequent questions utilize the quality-filtered city list.

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

4.1. Motivation

The relationship between urban density and green space access remains contested in planning theory. Compact city advocates argue that density enables efficient green space provision through economies of scale, while critics contend that densification reduces per-capita green space availability. This exploratory question examines whether denser neighborhoods within European cities have

better or worse access to parks and tree canopy, revealing that the density-access relationship operates as a *continuum within cities* rather than a fixed pattern across urban Europe.

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.

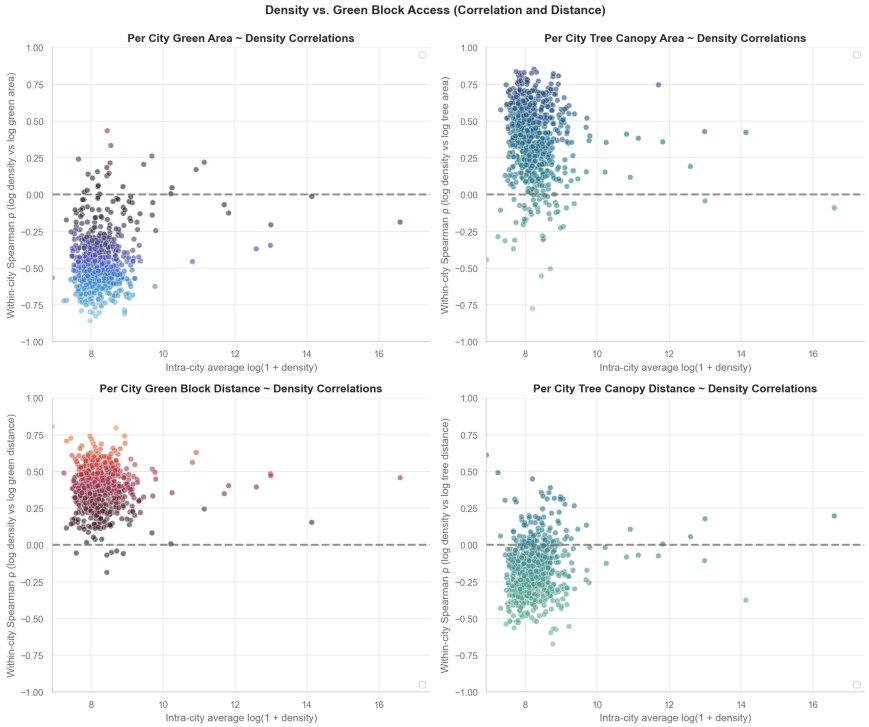

Figure 5: **Green space accessibility and tree canopy versus population density.** 2×2 grid comparing distance metrics (top row) and correlation analysis (bottom row) across 491 European cities. Left column: green blocks (parks). Right column: tree canopy. Points colored by Spearman correlation strength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

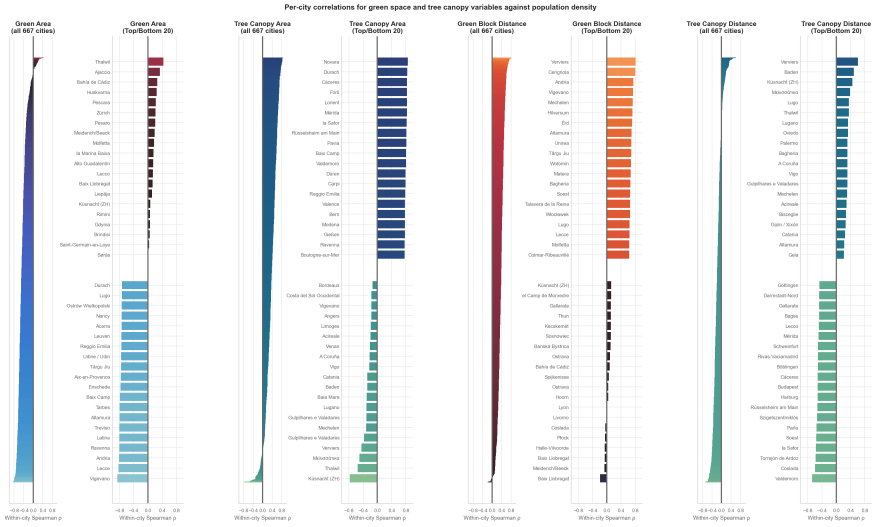

Figure 6: **Per-city correlation patterns for green space accessibility.** Diverging bar chart ranking cities by density-green correlation. For green blocks: only 4 cities show negative correlations (dense neighborhoods closer to parks); 487 cities show positive correlations (parks in peripheries). For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

4.5. Discussion

The near-universal positive correlation for green blocks within cities might initially suggest that densification inherently compromises green space access. However, **the critical finding is the absence of any systematic pattern across cities**. When plotting each city's mean density against either its density-green correlation or its mean distance to green space, no relationship emerges (Figure 5, bottom panels). Dense cities are not inherently worse for green access than sparse cities; a city averaging 5,000 persons/km^2 can provide equivalent or better absolute green access than a city at 1,000 persons/km^2.

This reveals that the within-city gradient reflects a **continuum of access within each urban area**, not a fundamental incompatibility between density and green space. Formal parks exhibit centrifugal placement—land scarcity in dense cores pushes large green spaces to urban peripheries—but this spatial logic operates at the intra-city scale. The magnitude of this gradient varies enormously across cities (ρ ranges from -0.06 to $+0.76$), demonstrating that planning decisions, not density per se, determine access equity.

The contrasting pattern for tree canopy reinforces this interpretation. Street tree programmes follow centripetal placement: they prioritise pedestrian-oriented cores where walking infrastructure already exists, inadvertently benefiting high-density residents. This suggests a scalable intervention: street tree programmes require minimal land acquisition and can be retrofitted into existing dense neighbourhoods, offering a pragmatic equity mechanism where park creation is politically or economically infeasible.

The planning implication is that density is not destiny. The within-city gradient reflects *how* green space is distributed relative to where people live, but cities can intensify while maintaining or improving access through deliberate intervention. Policy debates that assume densification inherently compromises environmental amenities conflate correlation with causation and ignore the decisive role of urban planning. Dutch cities like Spijkenisse achieve density *with* proximate parks through integrated planning, while peripheral park models (common in Belgium and parts of Eastern Europe) reproduce steeper

access gradients—but neither model correlates with overall city density.

For researchers, these findings suggest the importance of **multi-scale analysis**. Studying only aggregated city-level metrics would miss the within-city gradients that disadvantage dense neighbourhood residents. Conversely, observing only within-city patterns could wrongly suggest density is inherently problematic for green access, when the cross-city evidence shows no such relationship. Aggregate European statistics (e.g., "91% of nodes within 400m of green space") should not obscure the substantial within-city variation; planners in high-positive-correlation cities should prioritise infill parks or green corridor networks to reduce access gradients.

4.6. Extensions

Future work could incorporate green space quality metrics (size, facilities, maintenance); examine temporal changes as cities densify; conduct behavioural validation through mobility data or surveys; investigate how green space type (pocket parks vs. large regional parks) influences the density-access relationship; explore policy mechanisms that enable equitable access during densification; or develop scenario modeling to test impacts of proposed densification plans.

4.7. Reproducibility

Code: `paper_research/code/eg2_green_space/eg2_green_space.py`. Outputs: scatter plots, diverging bar chart visualization, and per-city correlation CSV in `outputs/` subfolder. Analysis restricted to cities with $\geq$100 nodes to ensure correlation stability.

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.

5.4. Results

Access to education is a tale of two Europes. In cities like Venezia (IT) and Almere (NL), over 60% of nodes are within a 5-minute walk of a school, and mean access distances are under 450m. By contrast, in places like Legionowo (PL) and Ludwigsburg (DE), mean distances exceed 700m and fewer than a third 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 2.6 (Almere, NL) to over 6 (Lugo, ES), and in the least equitable cities, nearly one in five nodes is severely underserved (Table 3).

Table 2: Best and worst access to education (mean distance and % within 400m).

City	Country	Mean Dist. (m)	% within 400m
Venezia	IT	312	72.6
Warszawa	PL	419	63.4
Almere	NL	433	61.7
Utrecht	NL	434	62.7
Lublin	PL	441	63.0
...			
Legionowo	PL	701	34.2
Harburg	DE	702	31.3
Wołomin	PL	703	32.0
Aschaffenburg	DE	704	32.9
Douai	FR	708	28.6

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

City	Country	P75/P25 Ratio	% Severely Underserved
Almere	NL	2.6	12.1
Tampere	FI	2.6	4.8
Västerås	SE	2.6	9.7
...			
Lugo	ES	6.2	18.3
A Coruña	ES	5.1	16.2
Focşani	RO	4.9	15.7

5.5. Discussion

Educational access varies dramatically across Europe, with some cities delivering walkable schooling for nearly all residents while others leave large swathes of children with long journeys. These patterns reflect planning legacies, urban form, and investment priorities.

5.6. Extensions

Future work could integrate school capacity data and enrolment boundaries to assess availability beyond proximity; examine correlations with socioeconomic characteristics; analyse temporal trends as school consolidation policies or demographic shifts alter demand; conduct comparative case studies of cities with exceptional access; model impacts of proposed new school locations; or validate network distance metrics with actual travel behaviour data (school bus routes, parent surveys).

5.7. Reproducibility

Code: `paper_research/code/eg3_education/eg3_education.py`. Aggregation pre-computes and caches per-city summary metrics to `temp/egs/eg3_education/education_city_data.parquet`. Only `Consistently Saturated` cities are included to ensure robustness of within-city equity measures.

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

6.1. Motivation

Network centrality metrics capture a location's structural importance within the street network, serving as proxies for pedestrian accessibility and potential footfall. High-centrality locations should theoretically attract more commercial activity. This question examines whether network centrality at multiple scales, combined with census demographics, can predict amenity distribution—specifically for eat & drink establishments and business & services. By training

predictive models on well-saturated cities, we can assess how consistently street network structure explains commercial location patterns across different urban contexts.

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 4), 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 5).

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 4: Best and worst predicted cities for eat & drink establishments (per-city R^2).

City	R^2	MAE	RMSE	Nodes
Alessandria (IT)	0.821	0.408	0.576	7,280
Maresme (ES)	0.815	0.407	0.598	8,539
Torino (IT)	0.807	0.423	0.588	110,122
Milano (IT)	0.766	0.445	0.593	387,236
Brescia (IT)	0.765	0.357	0.512	55,308
...				
Bergamo (IT)	0.588	0.451	0.618	65,130
Busto Arsizio (IT)	0.551	0.472	0.619	66,754
Heerlen (NL)	0.448	0.461	0.595	43,911
Gallarate (IT)	0.346	0.493	0.681	31,654

Table 5: Best and worst predicted cities for business & services establishments (per-city R^2).

City	R^2	MAE	RMSE	Nodes
Maresme (ES)	0.835	0.519	0.711	8,539
Cremona (IT)	0.810	0.487	0.660	7,784
Torino (IT)	0.801	0.559	0.759	110,122
Milano (IT)	0.760	0.586	0.771	387,236
Brescia (IT)	0.767	0.514	0.699	55,308
...				
Modena (IT)	0.611	0.648	0.879	18,207
Busto Arsizio (IT)	0.590	0.624	0.809	66,754
Treviso (IT)	0.595	0.696	0.881	13,430
Gallarate (IT)	0.398	0.656	0.906	31,654

Table 6: Top 10 features by importance for amenity prediction.

Eat & Drink		Business & Services	
Feature	Importance	Feature	Importance
Closeness 1,600m	0.122	Closeness 1,600m	0.149
Closeness 1,200m	0.119	Closeness 1,200m	0.146
Employment ratio	0.110	Population 15–64	0.099
Population density	0.109	Closeness 800m	0.098
Population 65+	0.106	Employment ratio	0.095
Population 15–64	0.098	Population density	0.090
Closeness 800m	0.095	Population 65+	0.085
Closeness 4,800m	0.067	Closeness 4,800m	0.079
Population <15	0.066	Population <15	0.053
Closeness 9,600m	0.057	Closeness 9,600m	0.045

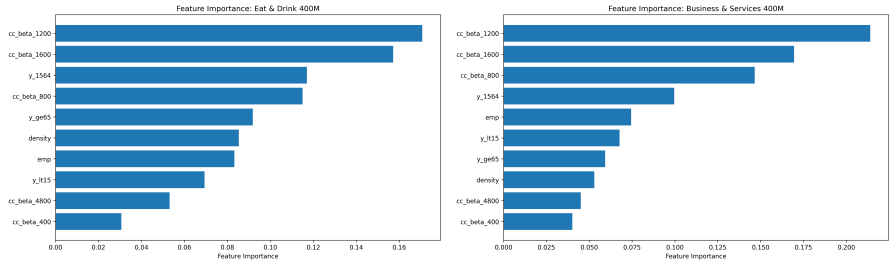

Figure 7: **Feature importance for amenity prediction models.** Left: Eat & drink. Right: Business & services. Intermediate-scale closeness centrality (1,200–1,600m) dominates both models, consistent with pedestrian catchment theory.

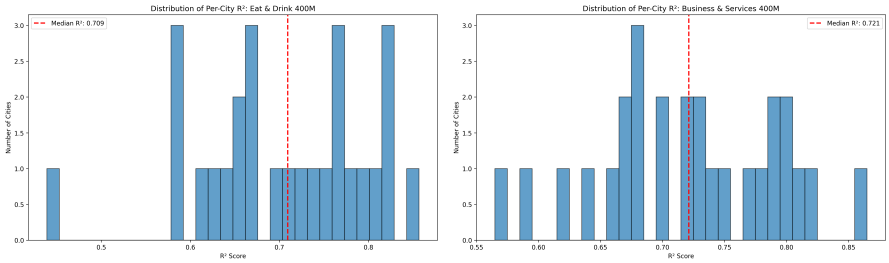

Figure 8: **Distribution of per-city R^2 scores.** Left: Eat & drink. Right: Business & services. Most cities cluster at high R^2 values, with a few outliers showing weaker network-amenity relationships.

6.5. Discussion

The strong predictive performance ($R^2 > 0.7$) demonstrates that street network structure, combined with basic demographics, explains a substantial share of amenity location variation. This validates the theoretical premise that commercial establishments optimise for pedestrian accessibility.

The dominance of intermediate-scale centrality (1,200–1,600m) aligns with urban economics theory: amenities serve neighbourhood catchments rather than immediate blocks or metropolitan regions. This scale corresponds roughly to 15–20 minute walking distances, resonating with "15-minute city" planning frameworks.

Why do some cities show weaker fit? Gallarate and Busto Arsizio—both satellite cities in the Milan metropolitan area—show lowest R^2 values. This may reflect: (1) amenity location driven by metropolitan-scale competition rather than local accessibility; (2) historical development patterns predating contemporary street networks; or (3) zoning constraints that override market logic. Heerlen (Netherlands), a former mining town undergoing economic restructuring, may exhibit legacy commercial patterns misaligned with current network structure.

Conversely, why do Italian cities consistently perform best? Several factors may contribute: (1) compact urban form with pedestrian-oriented development traditions; (2) strong correlation between network centrality and

historical commercial cores; (3) less car-dependent retail patterns than Northern European counterparts. The high fit for Milano (387,236 nodes, $R^2 = 0.77$) is particularly notable given its scale and complexity.

6.6. Extensions

Future work could incorporate additional explanatory variables (land values, zoning, competition indices); test whether residuals identify underserved areas or market opportunities; examine temporal stability of the centrality-amenity relationship; compare predictive performance across POI categories (e.g., education, health); or develop city-specific models to understand local deviations from pan-European patterns.

6.7. Reproducibility

Code: `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" (ville du quart d'heure) concept, popularised by Carlos Moreno and adopted in Paris's urban planning strategy, proposes that residents should access essential daily services—work, shopping, healthcare, education, leisure—within a 15-minute walk or cycle. This framework has gained traction as a sustainability and quality-of-life benchmark, yet systematic cross-city comparisons remain scarce. This question operationalises the 15-minute city concept using SOAR's pre-computed POI accessibility metrics, benchmarking European cities by the proportion of street network locations with complete access to all essential service categories.

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 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.

Bottom-performing cities 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 7: Top 10 cities by 15-minute city completeness (% nodes with access to all 10 POI categories within 1,200m).

City	Country	% Full Access	Mean Completeness
Venezia	IT	94.7	1.000
Baciu	RO	85.4	1.000
Modena	IT	85.3	1.000
Brașov	RO	85.2	1.000
Veliko Tarnovo	BG	84.6	1.000
Zamora	ES	84.4	1.000
Zagreb	HR	84.0	1.000
Roma	IT	83.6	1.000
Bolzano – Bozen	IT	83.4	1.000
Verona	IT	83.4	1.000

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

City	Country	% Full Access	Mean Completeness
Schweinfurt	DE	58.9	0.900
Como	IT	58.0	0.900
Wardenburg	DE	57.6	0.900
Lahti	FI	57.5	0.900
Roosendaal	NL	57.4	0.900
Brackel	DE	55.7	0.900
Iserlohn	DE	55.5	0.900
Aarau	CH	53.9	0.900
Uentrop	DE	53.9	0.900
Gulpilhares e Valadares	PT	48.7	0.900

Bottleneck categories reveal which services most frequently limit full 15-minute access. Table 9 ranks POI categories by mean access rate, identifying systematic gaps that planners could target for intervention.

Table 9: POI categories ranked by mean access rate within 1,200m (bottleneck analysis).

Category	Mean Access Rate (%)
Religious	87.2
Arts and Entertainment	92.4
Attractions and Activities	94.2
Education	97.2
Health and Medical	97.7
Public Services	97.9
Active Life	98.1
Eat and Drink	98.7
Retail	99.5
Business and Services	99.9

7.5. Discussion

The 15-minute city analysis reveals that complete local access remains rare across European cities. Several patterns emerge:

The gap between rhetoric and reality. While the 15-minute city has become a popular planning aspiration, few cities currently achieve universal access to all essential services within walking distance. Even cities with good average accessibility show substantial within-city variation, with peripheral neighbourhoods and lower-density areas falling short of the ideal.

Bottleneck categories matter. Full 15-minute access is often limited by one or two specific service categories rather than general accessibility deficits. Identifying these bottlenecks—whether arts & entertainment, health & medical, or other categories—provides actionable targets for infrastructure investment.

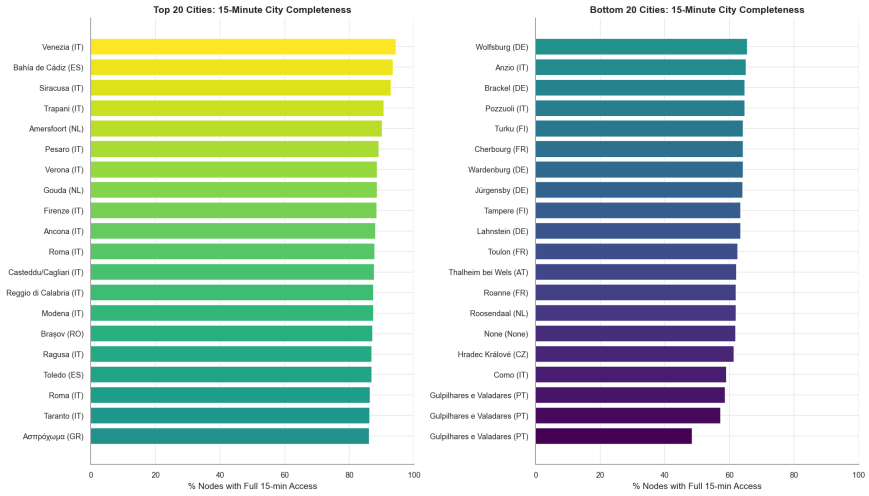

Figure 9: **15-minute city completeness ranking.** Left: Top 20 cities with highest proportion of nodes achieving full 15-minute access. Right: Bottom 20 cities. Colour intensity reflects completeness score.

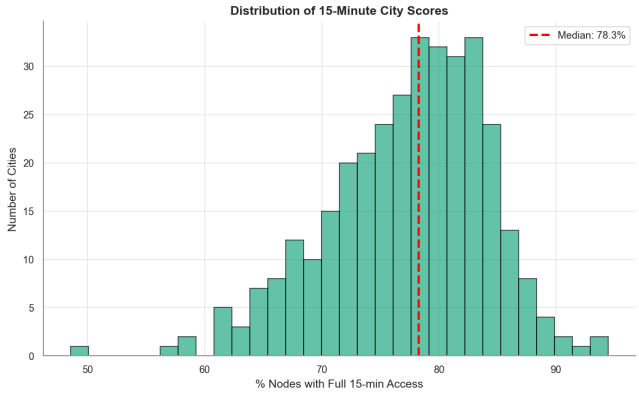

Figure 10: **Distribution of 15-minute city scores across European cities.** Histogram showing the percentage of nodes with full access to all 10 POI categories. Red dashed line indicates median.

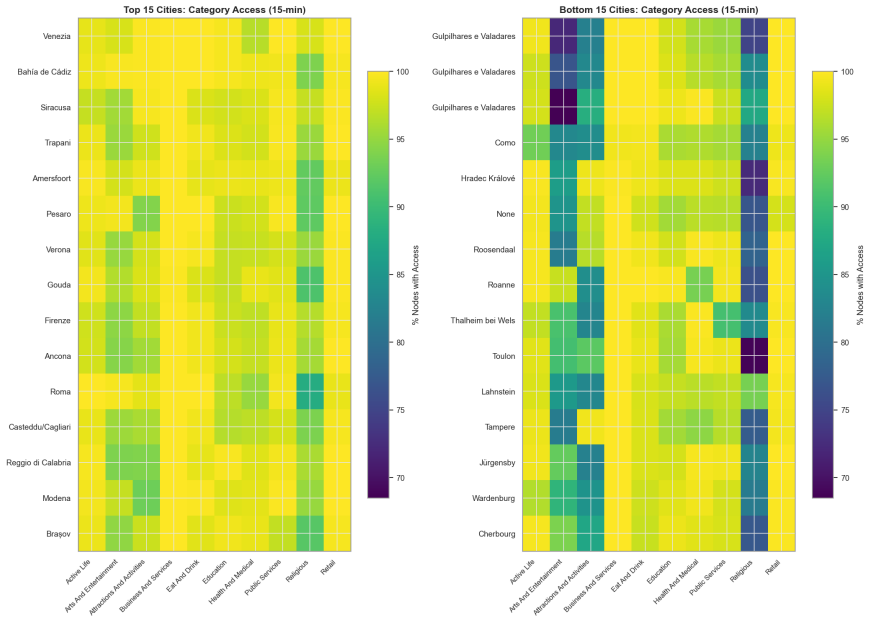

Figure 11: **Per-category access rates for top and bottom cities.** Heatmap showing which POI categories are most/least accessible within 1,200m. Columns represent categories; rows represent cities. Darker colours indicate lower access rates, revealing category-specific bottlenecks.

Urban form shapes completeness. Cities with compact, mixed-use development patterns consistently outperform sprawling or mono-functional urban areas. This reinforces the importance of integrated land-use and transport planning in achieving 15-minute city goals.

The metric is stringent but meaningful. Requiring access to *all* 10 categories sets a high bar that distinguishes genuinely walkable neighbourhoods from areas with partial accessibility. A location accessible to 9 of 10 categories may still require car trips for essential services, undermining the 15-minute city promise.

7.6. Extensions

Future work could weight POI categories by frequency of use (daily vs. weekly needs); incorporate travel time estimates from actual routing rather than network distance; examine how completeness varies by time of day (opening hours); compare 15-minute walking access with 15-minute cycling access; investigate correlations between completeness and socioeconomic characteristics; conduct temporal analysis as cities densify or decentralise; or develop scenario modelling to assess impacts of proposed new facilities.

7.7. Reproducibility

Code: `paper_research/code/eg5_poi_minutes/`. Analysis restricted to cities with more reliable combined POI saturation from Question 1.

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

8.1. Motivation

[TODO: Compare morphology distributions across cities.]

8.2. SOAR Metrics Utilized

[TODO: insert relevant morphology metrics]

8.3. Methodology

[TODO: Workflow description]

8.4. Results

[TODO: Key findings]

8.5. Implications

[TODO: Urban form insights]

9. Exploratory Question 7: What Are the Geographic Gradients in Amenity Access Across Europe?

9.1. Motivation

[TODO: Compute median walking distances to essential services per city; aggregate by country (where sufficient coverage) to reveal North-South and East-West gradients.]

9.2. SOAR Metrics Utilized

[TODO: Distance metrics for all POI categories, network distances]

9.3. Methodology

[TODO: Workflow description]

9.4. Results

[TODO: Key findings]

9.5. Implications

[TODO: Comparative planning insights]

10. Exploratory Question 8: How Can We Quantify Mixed-Use Development Across Cities?

10.1. Motivation

[TODO: Combine Hill diversity indices with building morphology metrics to classify neighborhoods along single-use ↔ mixed-use spectrum.]

10.2. SOAR Metrics Utilized

[TODO: Hill diversity (q=0,1,2), building FAR, coverage ratio, land-use accessibility]

10.3. Methodology

[TODO: Workflow description]

10.4. Results

[TODO: Key findings]

10.5. Implications

[TODO: Mixed-use planning insights]

11. Exploratory Question 9: Where Are the Best Opportunities for Strategic Densification?

11.1. Motivation

[TODO: Identify high-centrality, high-diversity nodes with low current population density as optimal densification targets.]

11.2. SOAR Metrics Utilized

[TODO: Beta-weighted closeness, Hill diversity, population density, building morphology]

11.3. Methodology

[TODO: Workflow description]

11.4. Results

[TODO: Key findings]

11.5. Implications

[TODO: Densification planning recommendations]

12. Exploratory Question 10: Which High-Demand Areas Lack Transit Infrastructure?

12.1. Motivation

[TODO: Map density against existing station locations to identify underserved high-demand areas.]

12.2. SOAR Metrics Utilized

[TODO: Centrality metrics, population density, existing transit infrastructure]

12.3. Methodology

[TODO: Workflow description]

12.4. Results

[TODO: Key findings]

12.5. Implications

[TODO: Transit planning recommendations]

13. Discussion

13.1. Cross-Cutting Themes

Across all ten 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.

13.2. 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.

13.3. 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

14. Conclusion

This paper presents ten exploratory questions demonstrating how integrated urban datasets like SOAR can address diverse research challenges in urban planning, geography, and data science. The contributions are:

1. **Breadth of applications**: Ten distinct analytical pathways—from data quality assessment to equity analysis, infrastructure gap identification, benchmarking, predictive modelling, and comparative geography—using multi-scale, node-level urban data.

2. **Data quality methodology**: The POI saturation analysis (Question 1) reveals systematic geographic patterns in crowdsourced data completeness, providing a methodological template for comparative analyses.

3. **Reproducible workflows**: Accessible code and clear methodological descriptions enable researchers with diverse backgrounds to engage with spatial network analysis and multi-scale accessibility metrics.

4. **Extension opportunities**: Each question identifies directions for more detailed, theoretically grounded investigations.

The standardised, multi-scale nature of SOAR facilitates comparative research across 699 European cities. As urban datasets continue to improve in coverage and quality, they offer growing opportunities for evidence-informed urban planning.

Acknowledgements

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

References

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