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] *Benchmark Urbanism, [TODO: Full postal address], Country*

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. This paper demonstrates applications of the SOAR (Scalable, Open, Automated, and Reproducible) urban data model [1] through worked examples that illustrate the dataset's contents, the types of questions it can address, 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 exploratory questions using SOAR to investigate diverse urban phenomena. Each question provides a reproducible workflow. The questions cover:

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

The questions are sequenced by analytical complexity. Question 1 assesses POI data quality and identifies cities with reliable coverage. Subsequent questions 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 re-

lated work on urban data applications; Sections 3–12 present the ten exploratory questions; 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. Exploratory Question 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$^{\beta}$), 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 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 (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 (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	$\mathbf{R^2}$	**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 shows population scale importance across categories, and Figure 2 presents model fit and z-score distributions. Figure 3 displays predicted vs. observed POI counts, while Figure 4 shows city quadrant classification across all POI categories.

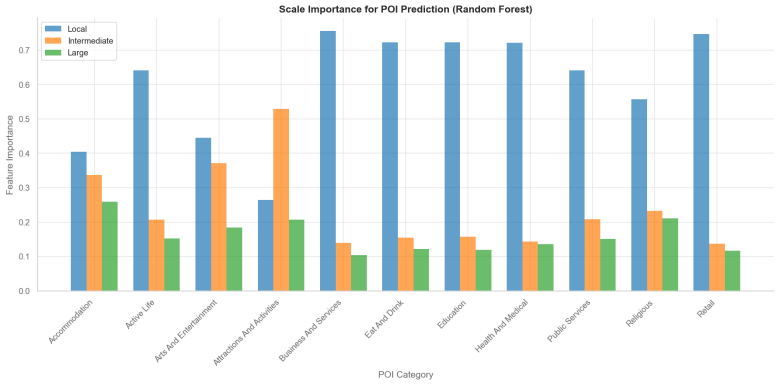

Figure 1: Feature importance analysis showing which population scale (local, intermediate, large) best predicts POI distribution for each category. Taller bars indicate stronger predictive power.

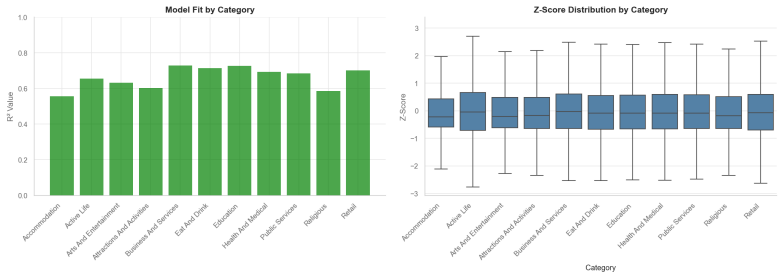

Figure 2: Exploratory data analysis. Left: Random Forest model fit (R^2) by POI category, with values ranging from 0.4–0.9 depending on category predictability. Right: distribution of z-scores across grid cells per category, revealing saturation patterns.

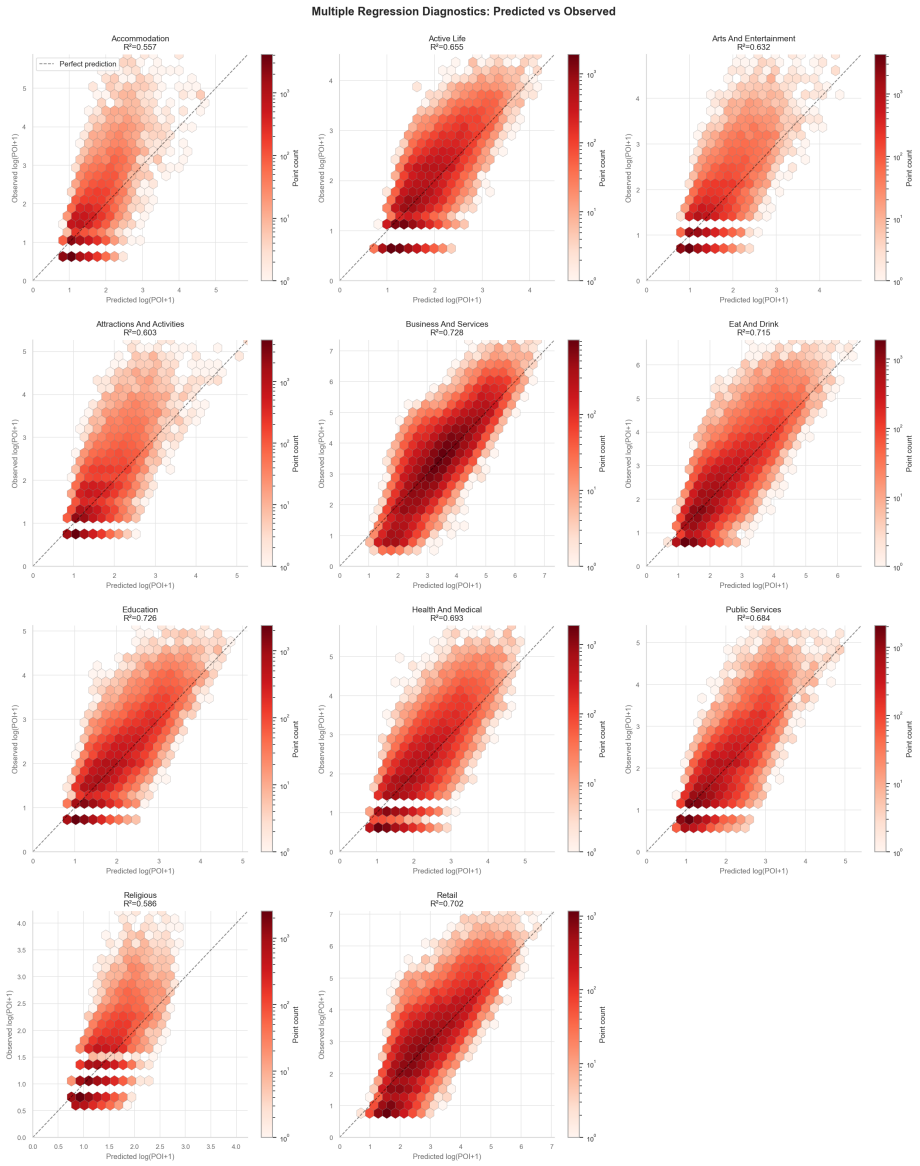

Figure 3: Regression diagnostics: predicted vs. observed POI counts (log scale) for each category. Points near the diagonal indicate accurate predictions; outliers represent grid cells with unexpected POI distributions.

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

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. Exploratory Question 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 $\mathbf{km^2}$ (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.

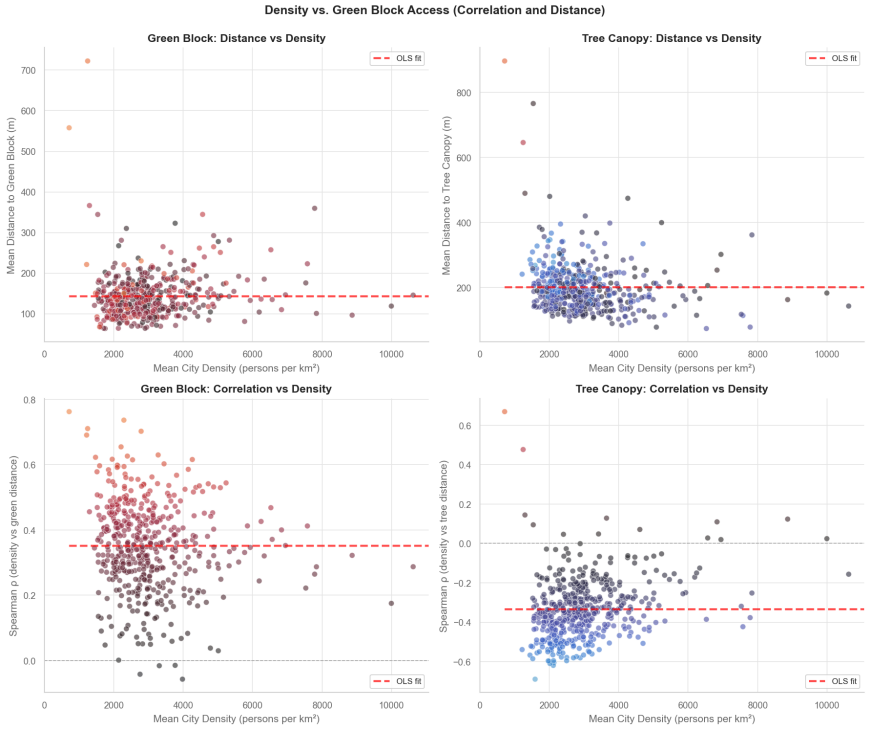

Figure 5: **Green space accessibility and tree canopy versus population density.** 2×2 grid comparing distance metrics (top row) and correlation analysis (bottom row) across 491 European cities. Left column: green blocks (parks). Right column: tree canopy. Points colored by Spearman correlation strength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

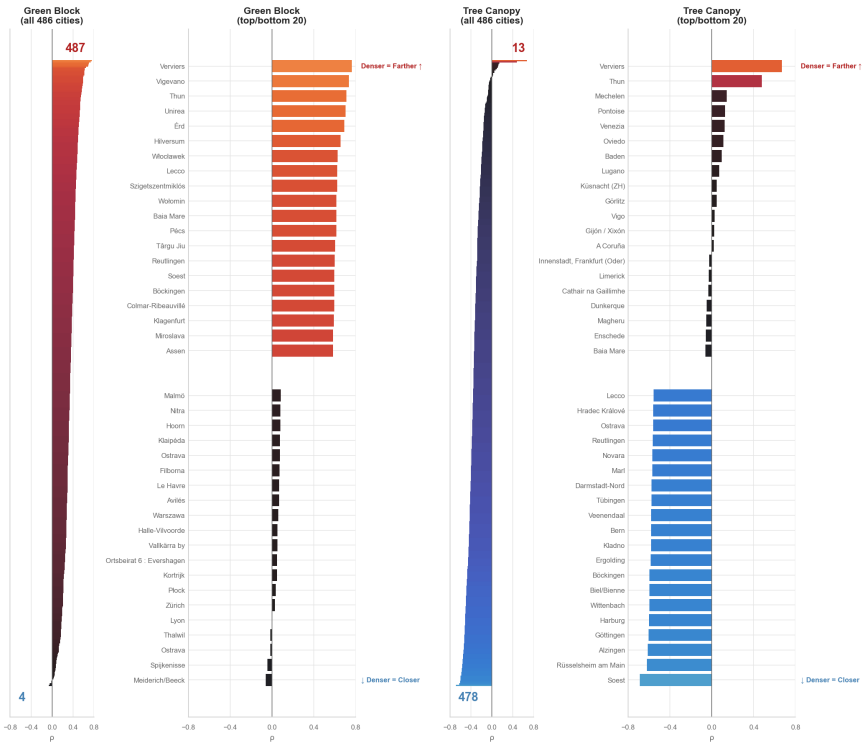

Figure 6: **Per-city correlation patterns for green space accessibility.** Diverging bar chart ranking cities by density-green correlation. For green blocks: only 4 cities show negative correlations (dense neighborhoods closer to parks); 487 cities show positive correlations (parks in peripheries). For 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. Exploratory Question 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. Exploratory Question 4: Can We Predict Undersupplied Neighborhoods for Amenities?

6.1. Motivation

[TODO: Train models using population density and network centrality to predict restaurant/cafe density; identify undersupplied neighborhoods. (Apply saturation filtering from Demonstrator 1 first.)]

6.2. SOAR Metrics Utilized

[TODO: Centrality metrics, population, POI counts, saturation z-scores]

6.3. Methodology

[TODO: Workflow description]

6.4. Results

[TODO: Key findings]

6.5. Implications

[TODO: Business and planning applications]

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

7.1. Motivation

[TODO: Identify streets with access to all 11 POI categories within 15-minute (1200m) and 20-minute (1600m) walks; rank cities by completeness. (Restrict to saturated cities from Demonstrator 1.)]

7.2. SOAR Metrics Utilized

[TODO: 1200m accessibility metrics, POI categories, population, saturation classification]

7.3. Methodology

[TODO: Workflow description]

7.4. Results

[TODO: Key findings]

7.5. Implications

[TODO: Policy implications]

8. Exploratory Question 6: How Do Density Patterns Vary Across European Cities?

8.1. Motivation

[TODO: Compare population density and building density morphology distributions across cities.]

8.2. SOAR Metrics Utilized

[TODO: Population density, network metrics, street segment characteristics]

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).