Utrecht Mobility Hub Analysis - Full Report

Geospatial Analysis of Transport Access and Infrastructure Gaps

Author: Manoela Calabresi

Date: April-May 2025

Tools: Python · GeoPandas · QGIS · HDBSCAN · UMAP

Background & Motivation

This project was originally developed as a capstone for a Data Analytics bootcamp. With a background in Architecture and Urban Planning, I designed the analysis to connect geospatial methods with strategic mobility planning. Using QGIS, Python, and clustering techniques, I explored where the mobility infrastructure in Utrecht could be most effectively expanded. After the bootcamp, I continued refining the methodology to better understand the full geospatial pipeline — from data to actionable insights.

Problem Statement

With mobility demand expected to grow by 25% by 2040, the Utrecht region faces an urgent need to optimize its network of mobility hubs. This study addresses the core question:

> Where should mobility hubs be implemented or upgraded in the Province of Utrecht to meet rising demand and ensure equitable, multimodal access?

Summary of Approach

This project analyzes the spatial dynamics of shared mobility in the Province of Utrecht to support strategic planning for future **mobility hubs**. With increasing pressure on urban transport systems, geospatial, demographic, and real-time mobility datasets were integrated to identify areas where additional infrastructure could have the greatest impact. The workflow combines **exploratory data analysis**, **feature engineering**, and **unsupervised clustering** techniques (DBSCAN, HDBSCAN), supported by **dimensionality reduction** using UMAP. Through this pipeline, accessibility patterns were assessed, mismatches between mobility supply and demand were detected, and high-priority zones for intervention were identified. The findings support **evidence-based planning** and help align transport investments with long-term urban development.

▼ Feature Engineering Summary

Feature	Description	Score Meaning / Usage
`hex_id`	Unique identifier for each hexagon	Indexing key

geometry	Geometry of the hexagon	Polygon for spatial operations
job_onsite	Count of onsite jobs	Raw job count
job_hybrid	Count of hybrid jobs	Raw job count
job_uncertain	Count of uncertain-location jobs	Raw job count
job_weighted	Weighted job count (based on remote feasibility)	job_onsite × 1.0 + hybrid × 0.6 + uncertain × 0.3
housing_density_utrecht_2025	Housing density projection for 2025	Higher = denser housing
absolute_growth_utrecht_2025	Projected housing growth (2015–2025)	Higher = faster growth
planned_housing_units	Future housing plans from provincial sources	Higher = more planned supply
planned_density_score	Combined score from growth + plans	Higher = more intense development
avg_vehicle_availability	Raw average of shared vehicles across time windows	Higher = more vehicles
capped_vehicle_availability	Same as above, but capped at 5	Controls for outlier distortion
log_vehicle_availability	log(1 + x) version for modeling	Reduces skew, used for clustering
has_ovfiets_access	1 if hex intersects OV-fiets station	Binary: 1 = yes, 0 = no
hub_distance_score	Score based on distance to nearest mobility hub	4 = ≤500m, 3 = 500–1000m, 2 = 1000–2000m, 1 = ≥2000m
hub_type_score	Score based on highest hub type in the hex	5 = Megastation, 4 = Intercity Station, 3 = Regional/Urban Hub, 2 = Local Station, 1 = P+R
hub_overall_score	hub_distance_score + hub_type_score	Range: 2 (low coverage) to 9 (close to major hub)
pt_line_distance	Distance to nearest PT line (bus/tram)	Continuous: Lower is better
pt_access_score	Score based on pt_line_distance	4 = ≤250m, 3 = 250–500m, 2 = 500–1000m, 1 = >1000m

* Revisiting Feature Engineering: Smoothing Skew and Sparsity

Initial HDBSCAN clustering attempts revealed distorted patterns, especially in dense urban areas. To improve spatial coherence and cluster stability, the feature engineering process was refined through targeted transformations:

- **Log Transformation** was applied to features with extreme skew (job_weighted, planned_housing_units, planned_density_score) to dampen outliers without losing relative magnitude.
- **Binary Flags** replaced sparse features with presence/absence indicators (e.g., future growth and density flags), enhancing interpretability and robustness.

Final Features Used for Clustering

Category	Feature	Description
Housing & Job Demand	planned_housing_units_log	Log-transformed count of future housing units
	job_weighted_log	Log-transformed job presence, weighted by relevance

	planned_density_score_log	Log-transformed score reflecting intensity of planned development
Urban Growth Signals (Flags)	absolute_growth_utrecht_2025_flag	Presence of projected absolute growth (2025)
	housing_density_utrecht_2025_flag	Flag for areas with increased housing density in the 2025 projection
Accessibility & Infrastructure	pt_access_score	Public transport accessibility score
	hub_overall_score	Composite distance score to nearest shared mobility hub
	log_vehicle_availability	Log-transformed count of shared vehicles nearby
	has_ovfiets_access	Binary indicator for OV-fiets availability

▼ Feature Engineering: Jobs

This section analyzes spatial job distribution in the Province of Utrecht by categorizing opportunities according to **work mode** (onsite, hybrid, uncertain) and aggregating them using a **hexagonal spatial grid**. The goal is to understand how different types of job opportunities align with urban mobility patterns and potential accessibility gaps.

1. Data Input and Structure

The core dataset consists of job postings geolocated across the region. Each posting was accompanied by a NUMPOINTS field representing how many job offers were aggregated at a given location. These postings were divided into three categories:

Job Mode	Description	Count (points)	Sum of NUMPOINTS
Onsite	Geolocated to a specific, physical address	9,228	179,390
Hybrid	Indicated some flexibility or dual presence	6,027	71,338
Uncertain	Ambiguous or generalized location data	8,417	194,759

To better understand the occupational landscape, jobs were classified using the **SBI sector codes**. Below is a simplified version of the classification table:

SBI Code	Category	Example Label
A	Agriculture	Onsite
C	Industry / Manufacturing	Onsite
F	Construction	Onsite
G	Retail / Trade	Onsite
H	Transport / Logistics	Onsite
I	Hospitality	Onsite
J	ICT / Tech	Hybrid
K	Finance / Insurance	Hybrid
M	Professional Services	Hybrid
N	Administrative Services	Hybrid
P	Education	Onsite
Q	Health / Social Work	Onsite

O	Government / Public Admin	Uncertain
S, T	Other	Uncertain

This classification helped anchor the job mode assumption with sectoral logic.

2. Spatial Aggregation

To better visualize spatial trends, all job locations were aggregated into a **uniform hexagonal grid** covering the entire province. The aggregation process included the following steps:

- **Spatial Join**: Each job location was matched with its corresponding hexagon based on geometry.
- **Summation of NUMPOINTS**: For each hex, the total number of job positions (`NUMPOINTS`) was calculated separately for each work mode category.
- **Handling missing data**: Hexes without jobs were assigned a value of zero to maintain uniformity across the dataset.

Job Counts Summary

Job Mode	Total Jobs	Number of Job Points	Average Jobs per Point
Onsite	67,518	9,228	~7.68
Hybrid	26,919	6,027	~4.63
Uncertain	73,278	8,417	~9.03

> Note: The “Total Jobs” column reflects the sum of the NUMPOINTS field, which estimates the number of job positions available at each location. A single job point may represent multiple roles, such as multiple vacancies within a hospital, office building, or industrial site.

3. Weighted Accessibility Model

To account for the **mobility demand** of each job type, weights were assigned based on how physically tied a job is to a specific place:

Job Mode	Weight	Rationale
Onsite	1.0	Must be accessed physically
Hybrid	0.6	Partial physical presence required
Uncertain	0.3	May be performed remotely

```
job_weighted = job_onsite * 1.0 + job_hybrid * 0.6 + job_uncertain * 0.3
```

This calculated field (`job_weighted`) serves as a proxy for **mobility-relevant job density**.

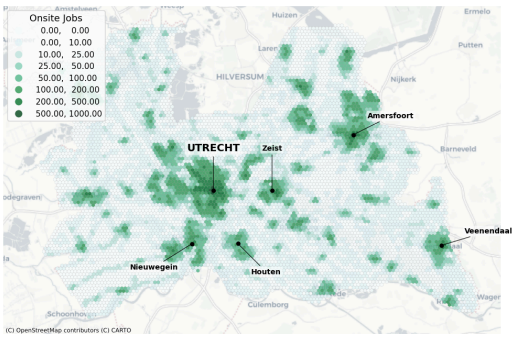
4. Final Results – Jobs per Hexagon

Category	Jobs (hex grid aggregated)
Onsite Jobs	67,518
Hybrid Jobs	26,919

Uncertain Jobs	73,278
Weighted Total	**~94,000**

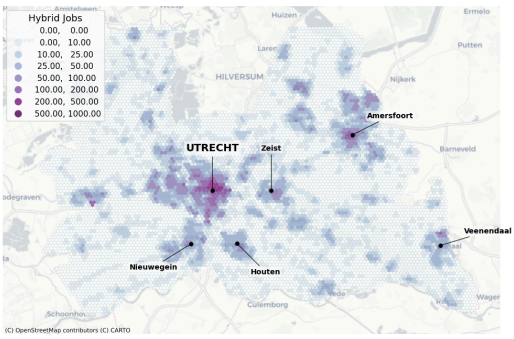
> These values match the total job counts computed directly from the point dataset, confirming that aggregation into the hex grid preserved overall job volume without duplication.

Onsite Jobs



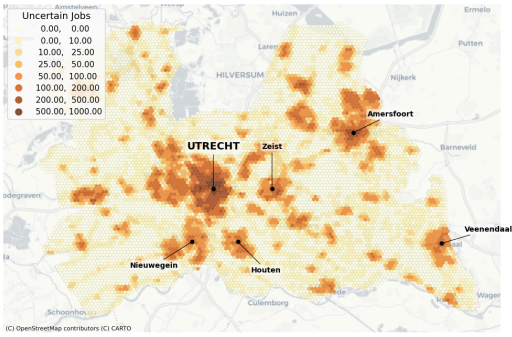

```
Summary Statistics:
Total hexagons     : 8,940
Non-zero hexagons  : 6,115 (68.4%)
Mean               : 20.1
Median             : 3.0
Max                : 835.0
```

Hybrid Jobs



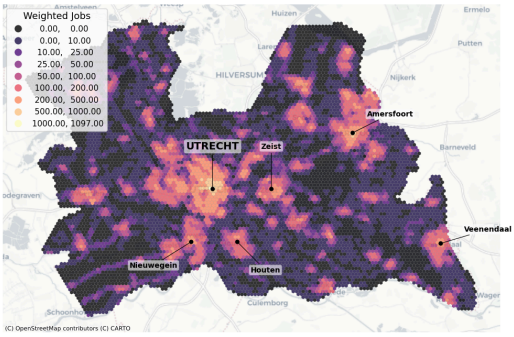

```
Summary Statistics:
Total hexagons     : 8,940
Non-zero hexagons  : 4,138 (46.3%)
Mean               : 8.0
Median             : 0.0
Max                : 256.0
```

Uncertain Jobs




```
Summary Statistics:
Total hexagons     : 8,940
Non-zero hexagons  : 5,471 (61.2%)
Mean               : 21.8
Median             : 2.0
Max                : 717.0
```

Weighted Jobs




```
Summary Statistics:
Total hexagons     : 8,940
Non-zero hexagons  : 6,197 (69.3%)
Mean               : 31.1
Median             : 3.0
Max                : 1097.0
```

Maps were generated for each layer using custom color palettes to visualize density gradients.

Takeaways

This is an **exploratory version** of a spatial job accessibility model. It should not be interpreted as a definitive representation of Utrecht's labor market but rather as a **prototype for integrating employment data with geospatial analysis**.

Potential Improvements

- Integrate **richer job data** (e.g. municipal APIs, job board APIs)
- Cross-reference **land use** or **building function datasets**

This first iteration fulfills its role as a **methodological proof of concept** — and a stepping stone for future geospatial labor analysis.

Data Sources

- **SBI Job Sector Codes**: Utrecht in Cijfers – Sectorindeling
- **Base Geometry**: CBS Wijken & Buurten, PDOK / Kadaster open data
- **Hex Grid**: Custom-generated in QGIS using Dutch projection (EPSG:28992)

▼ Feature Engineering: Housing

This section analyzes the spatial distribution of **housing demand** and **planned residential development** across the Province of Utrecht. The objective is to compare **historic housing growth trends** with **planned supply** as provided by municipal and provincial datasets.

---

1. Data Input and Structure

Two complementary datasets were used to represent housing conditions:

- **City-level data** (Utrecht municipality):
  - `housing_density_utrecht_2025`: Projected housing density in 2025
  - `absolute_growth_utrecht_2025`: Absolute unit growth from 2015 to 2025
- **Province-level data** (housing plans layer):
  - `planned_housing_units`: Total future units (aggregated across development types)
  - `planned_density_score`: Weighted score based on multi-family vs. single-family mix

---

2. Spatial Aggregation

To maintain consistency with other features, all housing data was aggregated using the **same hex grid** (250m) used for jobs and mobility. The steps included:

- **Spatial Join**: Each feature (e.g. projected housing density) was intersected with hex geometry
- **Summation or Averaging**: Values were aggregated per hex using `mean()` or `sum()` depending on layer type

---

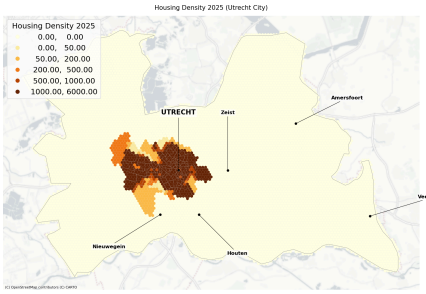
Unit Definitions for Housing Indicators

Indicator	Unit	Explanation
`housing_density_utrecht_2025`	Units per hectare	Projected housing density for 2025, based on Utrecht municipality urban forecasts

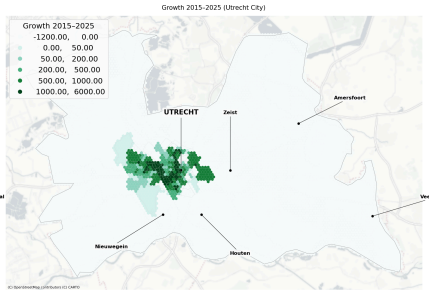
absolute_growth_utrecht_2025	New units	Total net housing units added between 2015–2025
planned_housing_units	Planned units	Future dwellings planned in provincial developments (aggregated count)
planned_density_score	—	Custom metric: 2×multifamily + 1×single-family; captures development intensity

3. Final Results – Housing Indicators per Hexagon

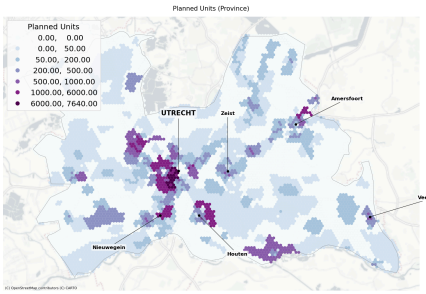
All four housing indicators use custom color palettes and consistent map styling. Fixed graduated bins normalize values for cross-map comparability.



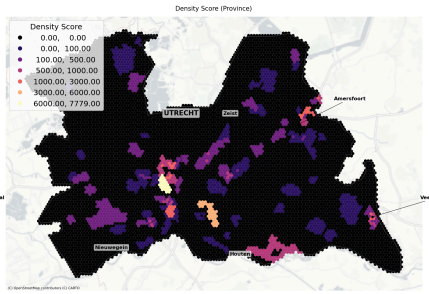
Summary Statistics:
Total hexagons : 8,940
Non-zero hexagons : 693 (7.8%)
Mean : 88.8
Median : 0.0
Max : 4286.0



Summary Statistics:
Total hexagons : 8,940
Non-zero hexagons : 552 (7.5%)
Mean : 29.4
Median : 0.0
Max : 5385.0



Summary Statistics:
Total hexagons : 8,940
Non-zero hexagons : 4,218 (47.2%)
Mean : 103.5
Median : 0.0
Max : 7640.0



Summary Statistics:
Total hexagons : 8,940
Non-zero hexagons : 1,417 (15.9%)
Mean : 70.7
Median : 0.0
Max : 7779.0

> Housing density, growth, and province-level planning across Utrecht Province (hex grid aggregation, graduated bins).

Takeaways

This analysis provides a clear comparison between **past urban growth** and **future residential development priorities**. Key observations:

- Planned housing supply is significantly more **geographically distributed** than recent growth.
- The **city center** dominates historical density, but provincial plans emphasize expansion to **eastern and southern nodes**.
- Some peripheral hexes score highly on **planned density**, making them promising targets for **multimodal transport integration**.

Data Sources

- **Housing Forecasts & Growth**: Utrecht in Cijfers – Woningvoorraad en Bouwprojecten – extracted via spatial layers used in official QGIS planning documents (2025 projections and 2015–2025 absolute growth).
- **Provincial Housing Plans**: Provincie Utrecht (housing_province.gpkg, updated 2024)
- **Hex Grid & Spatial Join Base**: Custom hexagon grid, EPSG:28992, intersected using GeoPandas

▼ Feature Engineering: Shared Mobility

This section analyzes the distribution of shared mobility services — including **bikes, scooters, cars**, and **OV-fiets public bikes** — across the Province of Utrecht. The goal is to create a **vehicle availability indicator** per hexagon, usable as a feature in clustering models and in accessibility reporting.

---

1. Data Input and Structure

The dataset combines multiple shared mobility layers, each representing a different **mode** (e.g. car, bike, scooter) and **time window** (morning and evening peak hours). All records are stored as **point geometries**, with some layers including an `available` flag indicating whether a vehicle was active at that timestamp.

The data was collected during **three different weekdays** to reflect variability in supply. For each day, vehicles were captured throughout the day and later grouped into two **rush hour intervals**:

- **Morning**: 06:30–10:00
- **Evening**: 16:00–20:00

This resulted in **six layers total** (2 time windows × 3 days), which were then concatenated to form a unified availability dataset. Only points with valid geometries were kept, and duplicates were removed.

Additionally, **NS OV-fiets station locations** were included as a separate binary access layer based on **500m buffer zones**.

Dataset	Description
Shared Cars (Morning)	06:30–10:00
Shared Cars (Evening)	16:00–20:00
Bikes & Scooters (Morning)	06:30–10:00
Bikes & Scooters (Evening)	16:00–20:00
OV-fiets Access	500m buffer around public bike stations

All datasets were cleaned and merged into a single layer named `vehicles_all`. Duplicate geometries and nulls were dropped, and missing `available` values were set to `1`.

---

2. Spatial Aggregation

To generate a vehicle availability feature per hex:

- **Spatial Join**: Points were assigned to hexagons using an `intersects` operation.
- **Averaging**: For each hex, the **average number of vehicles** across all four layers was computed.
- **OV-fiets Access**: Hexes within **500 meters** of an NS bike station were flagged with binary access (`ov_fiets_access = 1`).
- **Handling Missing Values**: Hexes without any available vehicles were assigned a value of zero.

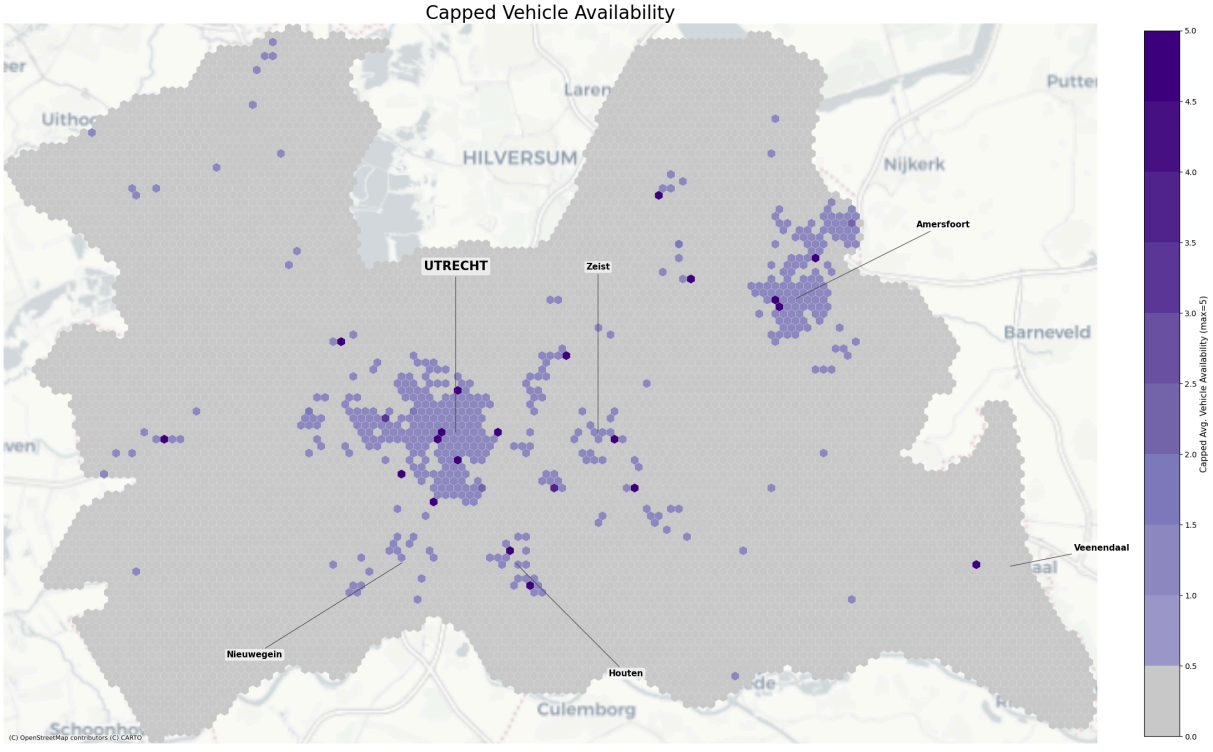
Three versions of the dynamic availability metric were produced:

Feature Name	Description
`avg_vehicle_availability`	Raw average across 4 layers
`avg_vehicle_availability_capped`	Same as above, but capped at 5 to reduce outlier distortion
`log_vehicle_availability`	`log(1 + x)` transformation to handle skewed distribution
`ov_fiets_access`	Binary indicator: 1 if hex is within 500m of a station

Vehicle Availability Summary

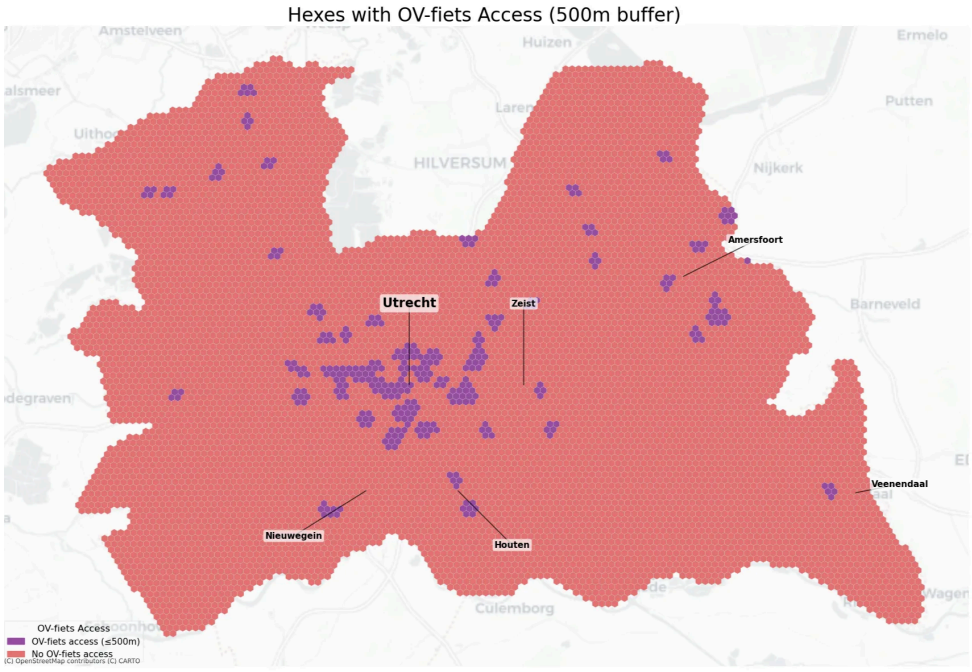
Feature	Sum	Mean	Max	Non-Zero Hexes
Avg. Vehicle Availability	1,074	0.12	345.6	330
Capped Avg. Availability	603	0.07	5.00	330
Log(1 + Availability)	1,639	0.18	5.85	330
OV-fiets Access (binary)	—	—	—	**243**

Capped Vehicle Availability



> ℹ️ Most hexagons have no recorded shared vehicle availability.

> OV-fiets expands spatial coverage, especially near train stations and smaller towns.



---

3. Feature Selection for Modeling

To integrate this feature into clustering:

- `log_vehicle_availability` is recommended for **modeling** (HDBSCAN) due to its continuous scale and reduced skew.
- `avg_vehicle_availability_capped` is recommended for **visualization** due to its interpretability and reduced distortion from high outliers.
- `ov_fiets_access` serves as a **binary accessibility flag**, complementary to vehicle availability.

All features were added to the master dataset (`hex_all_features`) for downstream use.

Takeaways

This is a **first attempt at quantifying shared vehicle accessibility** using raw availability and public bike access. It highlights **service clustering** in dense areas and supports broader evaluation of **multimodal coverage** across the region.

Data Sources

- **CROW-KpVV**: Shared Mobility Inventory – car-sharing, scooter data from different private providors
- **NS API + Provincie Utrecht Open Data**: OV-fiets station locations
- **Base Geometry**: CBS Wijken & Buurten, PDOK / Kadaster open data
- **Hex Grid**: Custom-generated in QGIS (EPSG:28992)
- **Basemap**: CartoDB Positron via Contextily

▼ Feature Engineering: Existing Mobility Hubs

This section focuses on quantifying **mobility hub accessibility** in the Province of Utrecht. It defines a **hub score per hexagon**, integrating two key dimensions:

1. **Proximity to hubs** (distance)
2. **Hub type / intensity** (hierarchical importance)

The resulting composite indicator is used to support clustering models and accessibility evaluations.

---

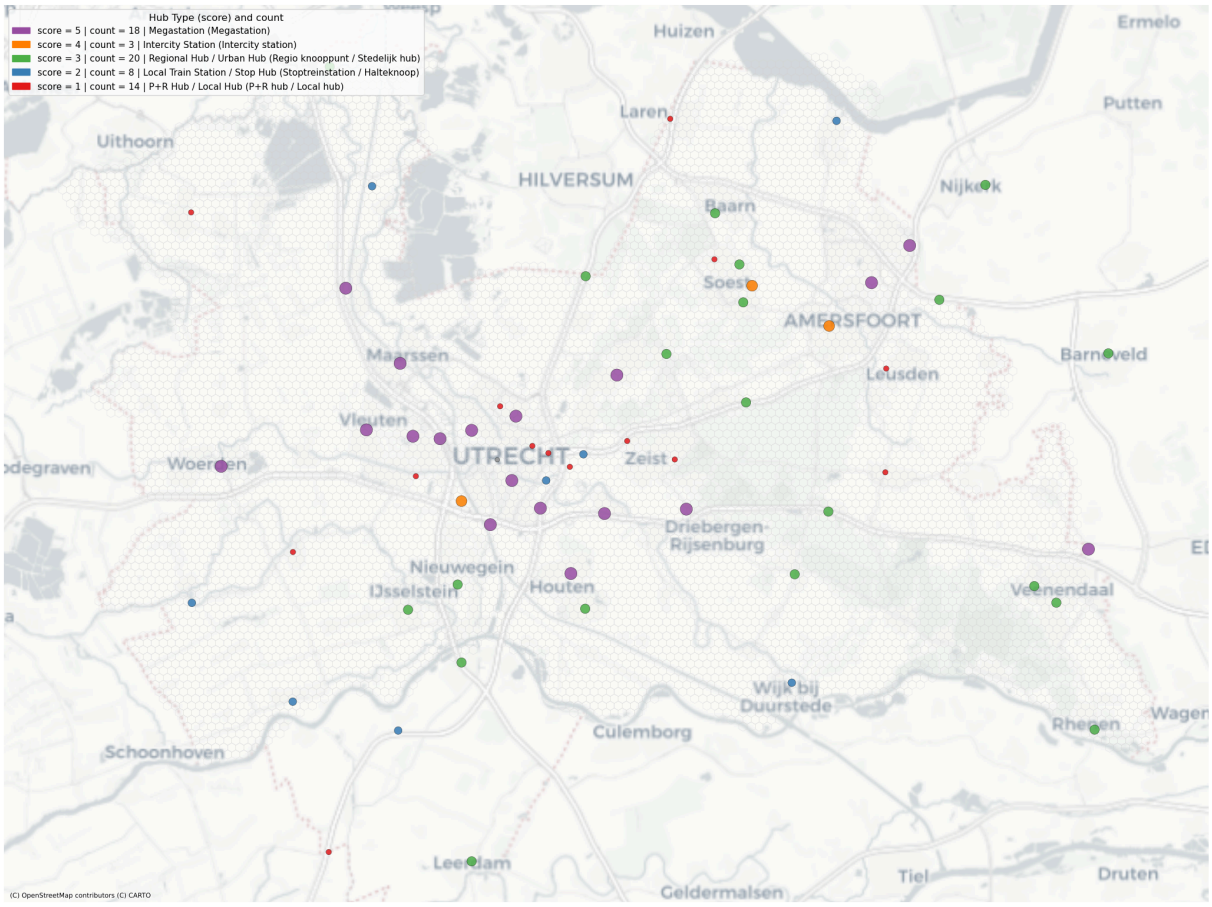
1. Hub Typology & Scoring

Each hub in the dataset was assigned a **type score** (1–5) based on its role in the public transport network:

Score	Hub Type	Dutch Name	Description
5	Megastation	Megastation	Major national/international station with extensive connections.
4	Intercity Station	Intercitystation	Key regional station serving intercity and local trains.
3	Regional Hub / Urban Hub	Regionaal knooppunt / Stedelijke hub	Medium-scale hub connecting regional/urban transit modes.
2	Local Train Station / Stop Hub	Stoptreinstation / Halteknoop	Small station serving local or commuter trains.
1	P+R Hub / Local Hub	P+R hub / Lokale hub	Local facility connecting parking, bus, or tram services.

> Each hub is tagged with only one type and receives the corresponding `hub_type`.

Mobility Hubs by Hierarchy (Score)



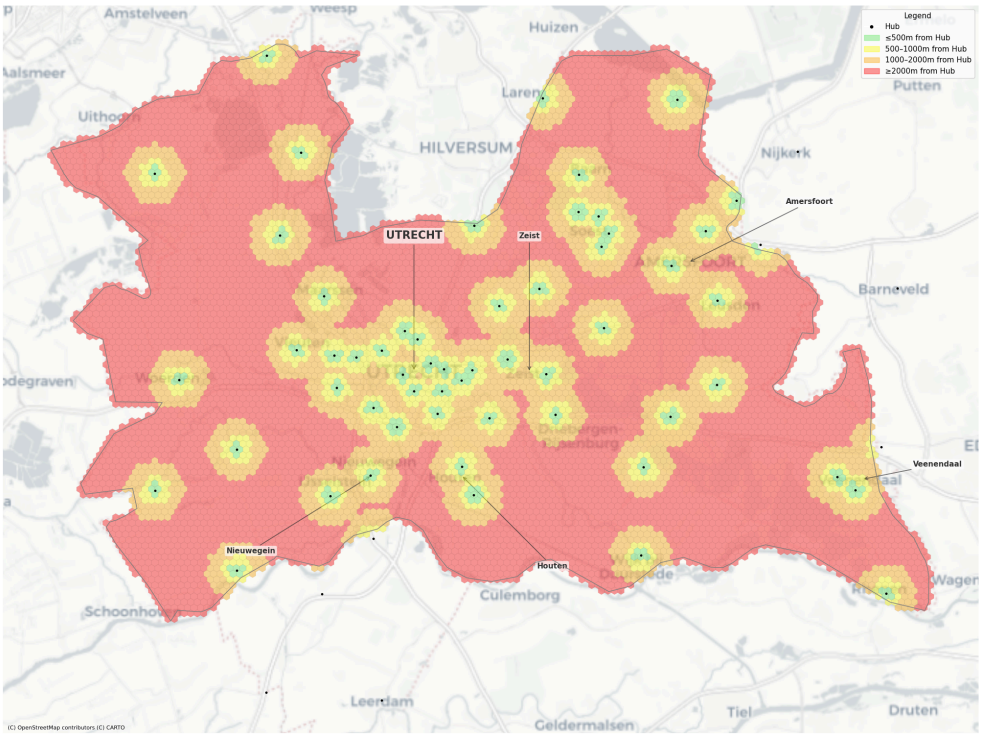
2. Buffer Zones (Distance Score)

Each hub was used to compute a **buffer zone** defining proximity ranges:

- ≤ **500m** → Distance Score: 4
- ≤ **1000m** → Distance Score: 3
- ≤ **2000m** → Distance Score: 2
- **2000m** → Distance Score: 1

> Each hexagon was assigned to the **nearest buffer category** based on its centroid and tagged with `hub_distance_score`.

Mobility Hubs by Distance Category (500m, 1000m, 2000m+)
with Utrecht Province Boundary and Main Cities



3. Feature Integration

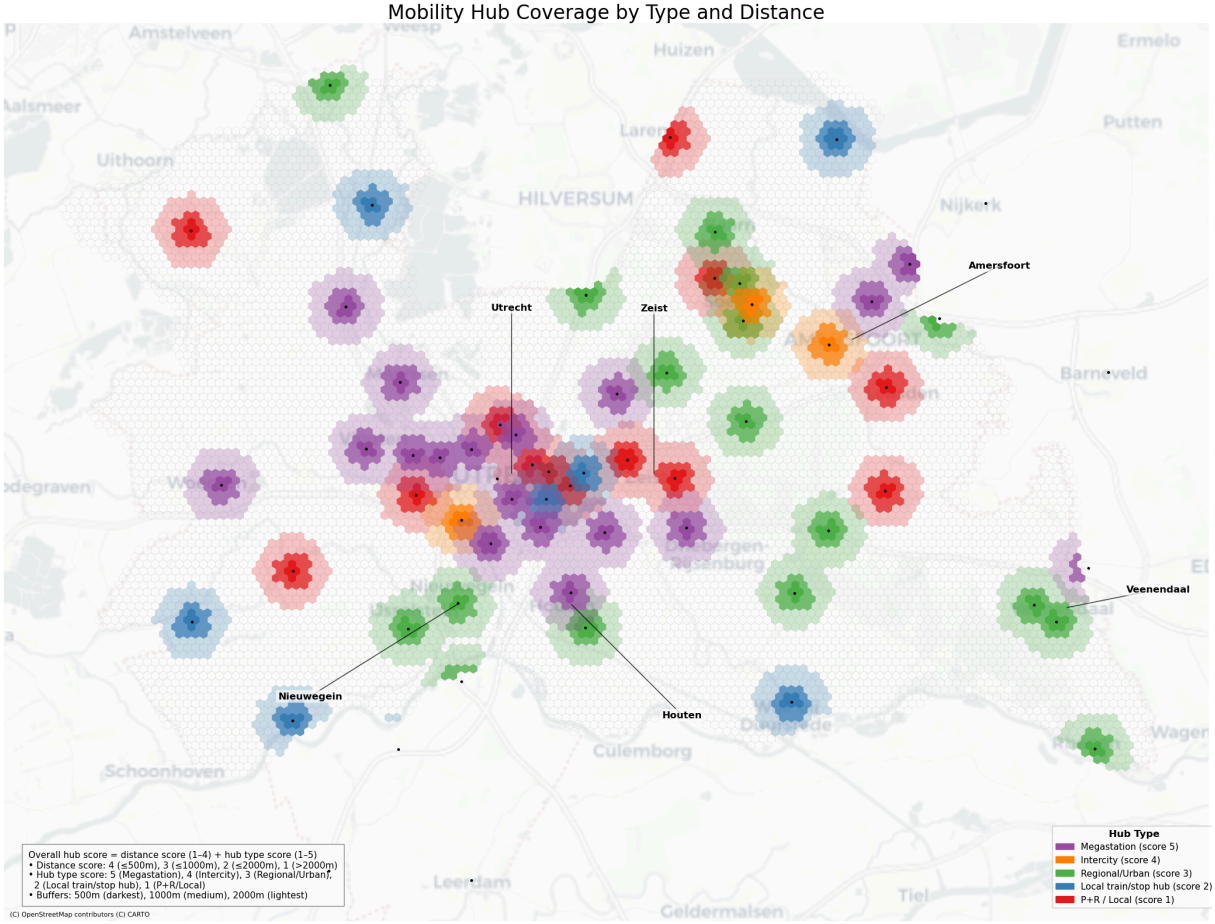
To compute the **final hub_overall_score**, both scores were combined:

```
hub_overall_score = hub_distance_score + hub_type_score
```

This ensures that **proximity** to a **high-ranking hub** yields a **higher score**, while distant hexes near minor hubs receive lower values.

If a hex had one score but not the other (e.g. proximity but no hub_type), a minimum value of 1 was enforced for both to ensure score consistency.

Mobility Hub Coverage by Type and Distance



> 🔍 About 16% of all hexes are influenced by one or more hubs. High scores are concentrated in urban cores (e.g. Utrecht, Amersfoort), especially near megastations and intercity stations.

4. Use in Modeling and Visualization

- `hub_overall_score` was added to the `hex_all_features` dataset.
- This feature is **used directly in clustering** models (e.g. HDBSCAN) to reflect access and intensity of transport services.
- A dedicated visualization gradient shows score contributions per hub type and distance zone.

Takeaways

This hub indicator captures **multimodal accessibility potential** at the hex level. It complements shared mobility indicators and helps identify **well-connected areas** vs. **underserved peripheries**.

Potential Enhancements

- Add **hub usage data** (e.g. ridership, boarding counts)

- Differentiate hub roles in **transfer chains** (e.g. modal split)
- Factor in **planned hubs** or upgrades (policy scenarios)
- Weight hub types based on **regional mobility impact**

Data Sources

- **Mobility Hub Locations**: Combined from two sources:
  - **NS API**: Official data from Nederlandse Spoorwegen (NS) providing station locations and classifications.
  - **Geo-Point Utrecht Open Data**: Station and transit hub layers accessed via the Provincie Utrecht open data portal.
- **Hub Typology Classification**: Manually assigned using fields like `FUNCTIEMNK`, `TYPE_TOEK`, and `TYPE_PU` from the datasets, cross-verified with NS classification schemes.
- **Base Geometry**: CBS Wijken & Buurten, PDOK / Kadaster open data
- **Hex Grid**: Custom-generated in QGIS using Dutch projection **Amersfoort / RD New (EPSG:28992)**.
- **Buffer Computation & Feature Engineering**: Performed using **GeoPandas** and **Shapely** in Python.

▼ Feature Engineering: Public Transport (PT) Networks

This section analyzes the **proximity of each hexagon to the nearest public transport line** (bus, tram, or train) across the Province of Utrecht. The result is a Public Tranport - **PT access score**, indicating how well each area is connected to the public transport network.

---

1. Data Input and Structure

The dataset includes the full **public transport network** in the Province of Utrecht, as provided by the **Provincie Utrecht Open Data Portal** (geo-point.provincie-utrecht.nl). It includes **multi-modal transit lines**, covering:

Mode	Dutch Term	Description
Train	*Trein*	Intercity and local/regional rail services
Tram	*Tram*	Light rail services (e.g. Utrecht tram lines)
Bus	*Bus*	Local and regional bus lines
Express Bus	*Snelbus* or *Interliner*	High-speed regional connections
Neighborhood Bus	*Buurtbus*	Small-scale, low-frequency routes for rural areas

The data is structured as **line geometries**, each tagged with metadata such as service type (`OV_TYPE`), operating company (`OV_MAATSCH`), and capacity indicators.

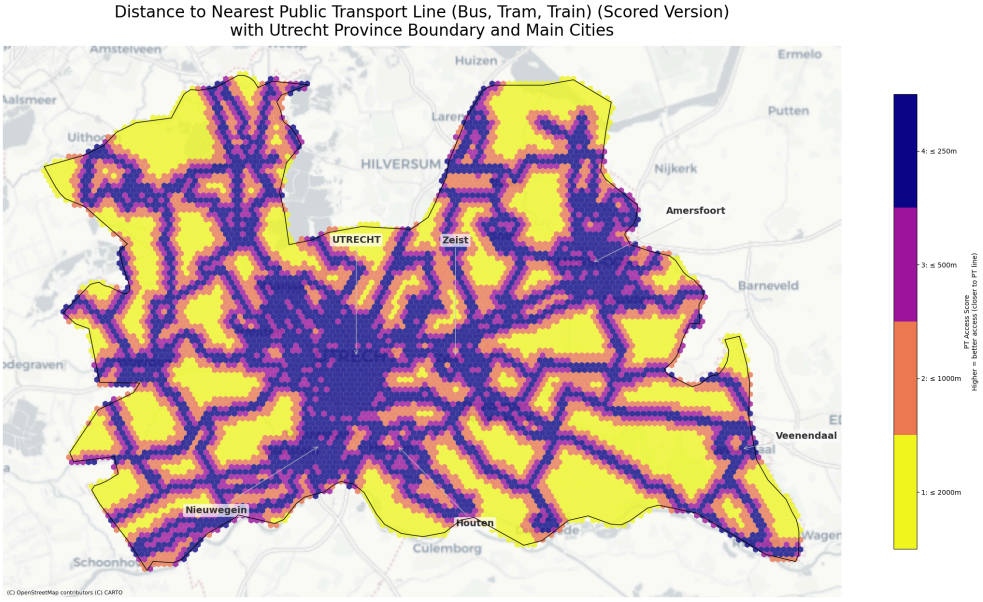
---

2. Scoring Methodology

To create a hex-level access score:

- **Centroid Calculation**: Each hexagon's geometric center was computed.
- **Distance Measurement**: The distance from each centroid to the **nearest public transport line** was calculated.
- **Scoring**: Based on distance, hexagons were assigned a categorical **PT access score** from 1 to 4.

Access Score	Distance to PT Line	Interpretation
4	≤ 250 meters	Excellent access (direct proximity)
3	≤ 500 meters	Good access
2	≤ 1000 meters	Moderate access
1	≤ 2000 meters	Limited access



> Note: Higher scores reflect better access (shorter walking distance to PT infrastructure).

---

3. Use in Modeling

This categorical score (`pt_access_score`) was added to `hex_all_features` as a **spatial accessibility variable**. It can be used to:

- Weight clustering around existing PT corridors
- Compare shared mobility coverage vs. PT access
- Identify regions with **low PT access but high housing growth**

Takeaways

- PT access is **strongest around Utrecht, Amersfoort**, and regional rail corridors.
- Peripheral rural zones often score **1 or 2**, indicating the need for better multimodal integration.
- This feature helps reveal spatial equity gaps in infrastructure.

Data Sources

- **Public Transport Lines**: Provincie Utrecht Open Data Portal
- **Hex Grid**: Custom-generated in QGIS (EPSG:28992)

- **Base Geometry**: CBS Wijken & Buurten, PDOK / Kadaster

▼ UMAP + HDBSCAN Clustering: Setup and Interpretation

With the feature set properly transformed and cleaned, dimensionality reduction and clustering were conducted using UMAP followed by HDBSCAN. This two-step pipeline enabled the uncovering of latent spatial structures and the segmentation of the study area into interpretable zones.

Methodology

1. Scaling the Input Data

Before clustering, all selected features were scaled using `RobustScaler`. This method was chosen to reduce the influence of outliers while preserving the structure of the majority of values.

2. Dimensionality Reduction with UMAP

UMAP (Uniform Manifold Approximation and Projection) was used to project the high-dimensional feature space into two dimensions. Key parameters:

- `n_neighbors` (default): Controls local vs global structure (kept default for general balance)
- `init="random"`: Ensures varied initializations, reducing deterministic artifacts
- `random_state=42`: For reproducibility

The UMAP output captured relationships between hexes based on all selected features, creating a spatially-aware embedding.

3. Clustering with HDBSCAN

HDBSCAN was applied to the UMAP embedding to identify natural groupings. Parameters:

- `min_cluster_size=15`: Minimum number of hexes to form a cluster
- `min_samples=5`: Controls density sensitivity and noise tolerance

This step allowed for automatic detection of both dense and sparse zones, while assigning some areas as noise when they did not fit any cluster.

Output and Visual Interpretation

Each clustering run generated:

- A new column `umap_cluster` with cluster labels
- A second column `cluster_name`, with formatted labels (e.g., `C1 (108)` for Cluster 1 with 108 hexes)
- A color-coded map with clearly separated regions and multi-column legends for readability

In addition, summary tables were produced with mean values for each feature within each cluster and the number of hexes per group. These summaries supported the interpretation of spatial trends and assisted in labeling the clusters based on their dominant characteristics.

▼ Thematic Clustering: Housing and Jobs –Urban Demand Set

This feature set emphasizes **areas with high combined pressure from housing development and job activity**, while controlling for current mobility infrastructure. It is designed to reveal zones where demand is rising but mobility access may lag behind.

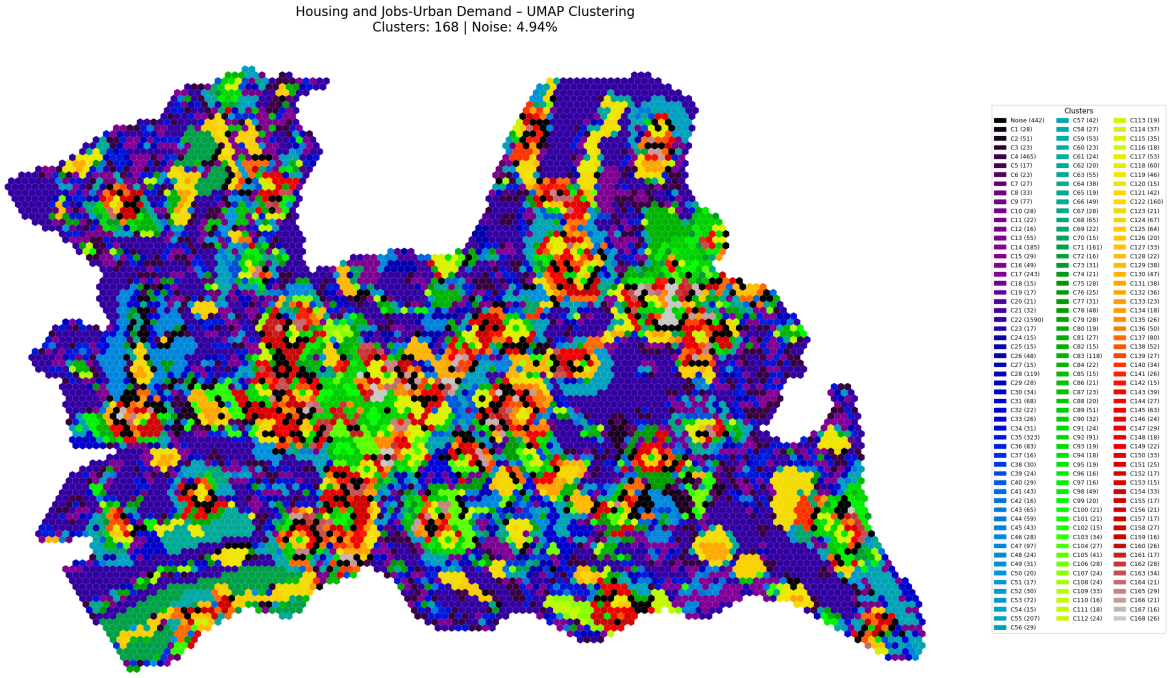
Features included:

- job_weighted_log
- planned_housing_units_log
- hub_overall_score

This combination offers a strategic view of urban development hotspots relative to shared mobility hub coverage.

Cluster Overview

Clusters are colored based on shared demand/access profiles. Areas in black are classified as noise (outliers).



Set's Key Spatial Patterns

1. **Well-served urban centers** emerge as distinct clusters, particularly in Utrecht, Amersfoort, and their adjacent nodes. These zones combine high job density, strong planned housing, and robust hub access.
2. **Peripheral development areas**, especially in the south and northeast, form clusters with **strong demand signals** but lower hub access, suggesting a mismatch between population growth and infrastructure.
3. **Transitional areas**—at the fringe of high-density cores—cluster with moderate values, indicating potential for strategic investment in transport and land use integration.

Gap Analysis Methodology

To systematically identify underserved clusters, a custom gap score was computed using the following logic:

> gap_score = (job_weighted_log + planned_housing_units_log) / hub_overall_score

This gives higher priority to zones with **strong housing and employment growth signals** but **low hub accessibility**.

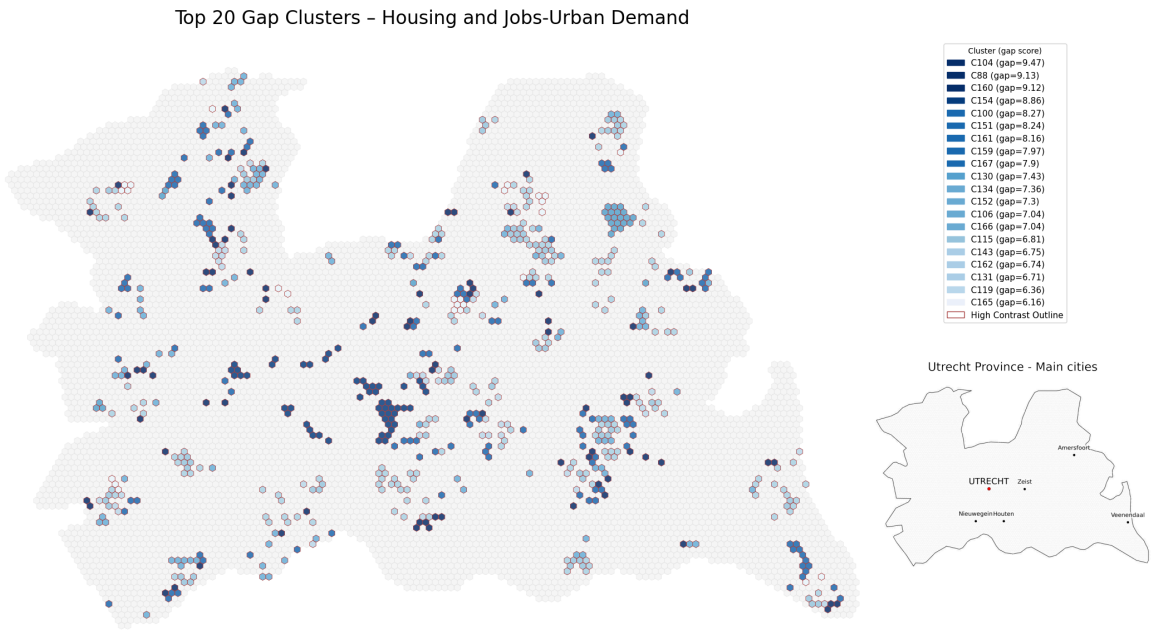
Top 20 Underserved Clusters in Housing and Jobs –Urban Demand Set

The table below ranks clusters by their calculated **gap score**, highlighting candidates for new hubs or service adjustments.

Cluster	UMAP Cluster	Hex Count	Gap Score	Hub Score	Avg. Jobs (raw)	Avg. Planned Units
C104	104	41	9.47	3.0	3,860.72	3,860.72
C88	88	51	9.13	4.0	3,635.29	3,635.29
C160	160	17	9.12	3.0	4,478.55	735.60
C154	154	17	8.86	3.0	2,648.59	1,419.53
C100	100	21	8.27	5.0	2,848.00	3,353.65
C151	151	17	8.24	4.0	4,459.12	788.22
C161	161	28	8.16	3.0	3,437.36	349.13
C159	159	26	7.97	3.0	2,406.14	617.57
C167	167	26	7.90	3.0	5,170.31	174.50
C130	130	38	7.43	2.0	1,072.77	123.81
C134	134	26	7.36	4.0	3,529.82	242.36
C152	152	15	7.30	4.0	1,705.73	709.26
C106	106	24	7.04	2.0	593.42	595.24
C166	166	16	7.04	3.0	2,713.77	134.66
C115	115	18	6.81	5.0	2,888.75	684.59
C143	143	27	6.75	4.0	2,525.13	355.58
C162	162	34	6.74	3.0	1,874.46	308.40
C131	131	36	6.71	2.0	2,176.14	65.41
C119	119	15	6.36	2.0	1,474.59	86.99
C165	165	21	6.16	3.0	2,127.35	111.91

Gap Cluster Map

This visualization isolates the **top 20 clusters with the highest gap scores**, overlaying their location on the province-wide hex grid. These zones represent **key intervention points** for shared mobility expansion.



Interpretation

- The clustering process helped isolate **zones where urban growth is outpacing infrastructure**.
- Several of the top clusters are in **growing peri-urban or suburban areas**, such as Veenendaal, Zeist, and southwestern Utrecht.
- Many of these areas have a high **planned housing load** but remain under-connected to major mobility hubs.
- These findings serve as **a decision-support layer** for prioritizing mobility hub siting, especially in areas forecasted for residential or employment growth.

▼ Thematic Clustering: First and Last Mile Need Zones Set

To identify areas where first and last mile connectivity may be insufficient, UMAP + HDBSCAN clustering was applied to a feature set that combines demand indicators (jobs, housing) with multiple forms of accessibility (PT access, distance to hub, vehicle availability).

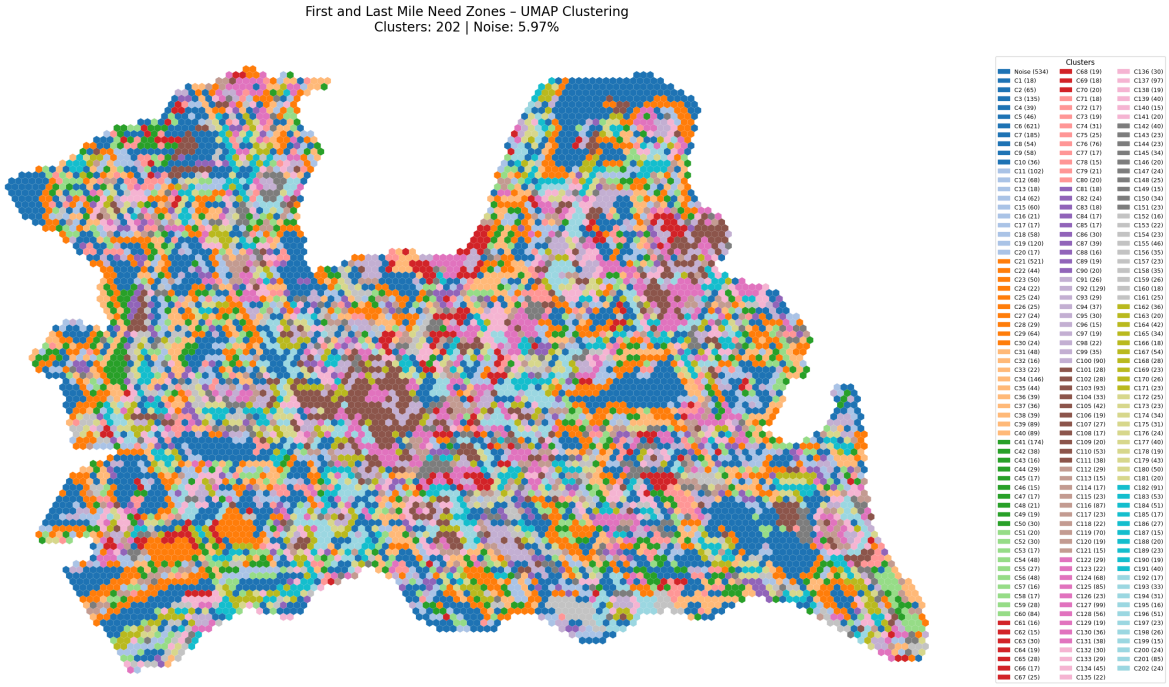
Features used in clustering:

- `planned_housing_units_log`
- `job_weighted_log`
- `pt_access_score`
- `hub_distance_score`
- `log_vehicle_availability`

Cluster Overview

- **Clusters identified**: 202

- **Noise points**: 5.97%
- **Clustering resolution**: High (due to mixed accessibility and demand indicators)
- **Fragmentation**: Expected given the inclusion of both continuous and categorical spatial conditions



Gap Analysis Methodology

To prioritize clusters for intervention, a **gap score** was computed to reflect areas where **demand is high** but **supply is weak**. Each indicator was preprocessed to align with this logic.

Scoring Logic

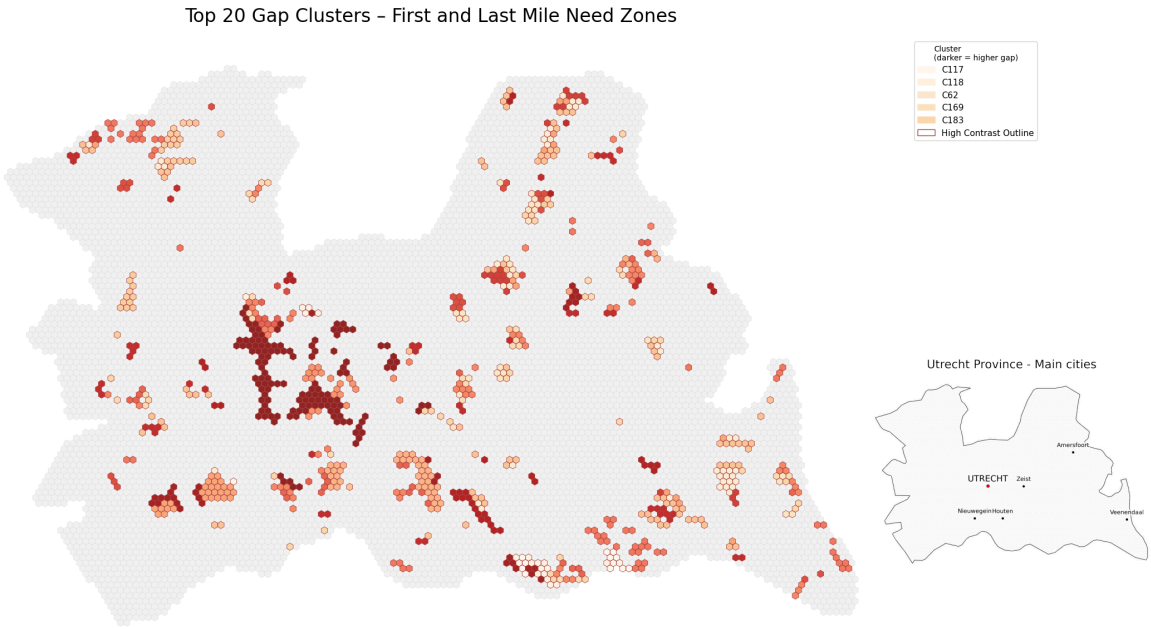
Feature	Weight	Interpretation
planned_housing_units_log	0.30	High value = high demand
job_weighted_log	0.30	High value = high demand
pt_access_score	0.15	Low value = low supply (inverted)
hub_distance_score	0.15	High value = low supply
log_vehicle_availability	0.10	Low value = low supply (inverted)

> To reflect a demand-supply mismatch, features were inverted when necessary to ensure that higher values consistently indicated higher need. All features were then normalized and combined using the weights above.

Top Gap Clusters: Highest Priority Zones

UMAP clusters were ranked based on a composite **gap score** derived from five key features. Scores reflect clusters where **demand is high** but **access is limited**. The map below highlights the 20 clusters with the

highest gap scores. Darker shades represent areas with significant demand but **insufficient access** to public transport or shared mobility.



Gap Cluster Table – First and Last Mile Need Zones

Cluster	Gap Score	Job Score	Planned Housing	Hub Score	PT Access	OV-fiets Access
C118	0.879	0.93	6.27	2.00	0.00	0.00
C119	0.842	2.53	6.39	2.00	1.71	0.00
C63	0.816	2.34	4.30	2.00	0.90	0.00
C170	0.805	4.40	4.30	3.00	1.85	0.00
C184	0.768	4.66	5.85	3.00	3.00	0.00
C183	0.764	3.26	5.09	3.00	1.97	0.00
C62	0.738	1.57	2.88	2.00	0.69	0.00
C172	0.731	2.38	4.19	2.00	2.00	0.00
C165	0.727	3.17	7.81	3.00	3.62	0.00
C83	0.723	0.28	5.01	2.00	1.00	0.00
C98	0.713	4.74	8.21	4.00	3.76	1.00
C199	0.710	4.93	5.96	3.00	4.00	0.00
C87	0.710	1.79	0.36	2.00	0.00	0.00
C129	0.709	4.81	7.08	4.00	3.91	0.00
C166	0.703	3.78	2.65	3.00	1.96	0.00
C102	0.695	0.00	4.30	2.00	0.83	0.00

C191	0.694	4.42	4.72	2.00	4.00	0.00
C48	0.693	1.70	2.09	2.00	1.00	0.00
C114	0.688	0.05	5.88	2.00	1.73	0.00
C104	0.685	3.90	5.89	3.00	3.51	0.00

Set's Key Spatial Patterns

1. **Darker clusters represent highest priority zones**, where rising demand is not matched by infrastructure supply.
2. These clusters are primarily located in **suburban areas or new development zones**, often beyond the reach of existing PT and mobility services.
3. None of the top 20 clusters have **OV-fiets access**, indicating potential for shared bike expansion.
4. Several clusters (e.g. C103, C113, C47) are spatially grouped, which strengthens the case for **area-based shared mobility interventions** rather than isolated pilots.

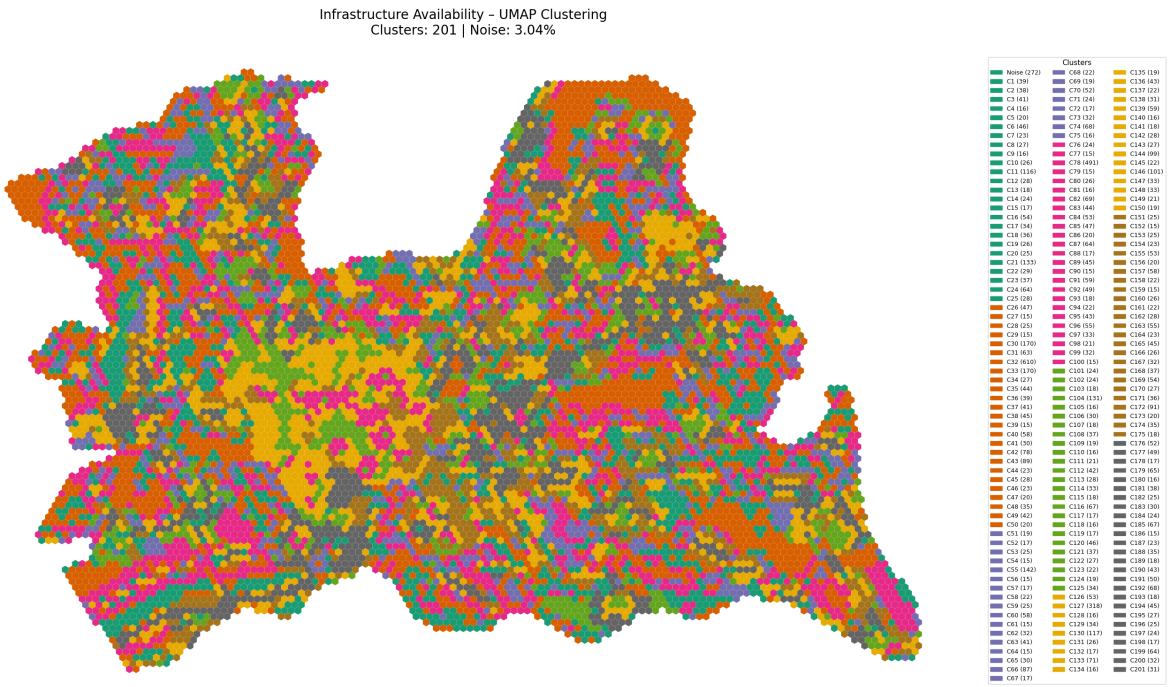
▼ Thematic Clustering: Infrastructure Availability Set

This feature set offers a system-level view of **existing infrastructure coverage** relative to projected demand. It highlights zones with high multimodal accessibility as well as spatial mismatches where infrastructure lags behind future urban growth.

Included features:

- `pt_access_score`
- `hub_overall_score`
- `has_ovfiets_access`
- `planned_housing_units_log`
- `housing_density_utrecht_2025_flag`
- `job_weighted_log`

Cluster Overview

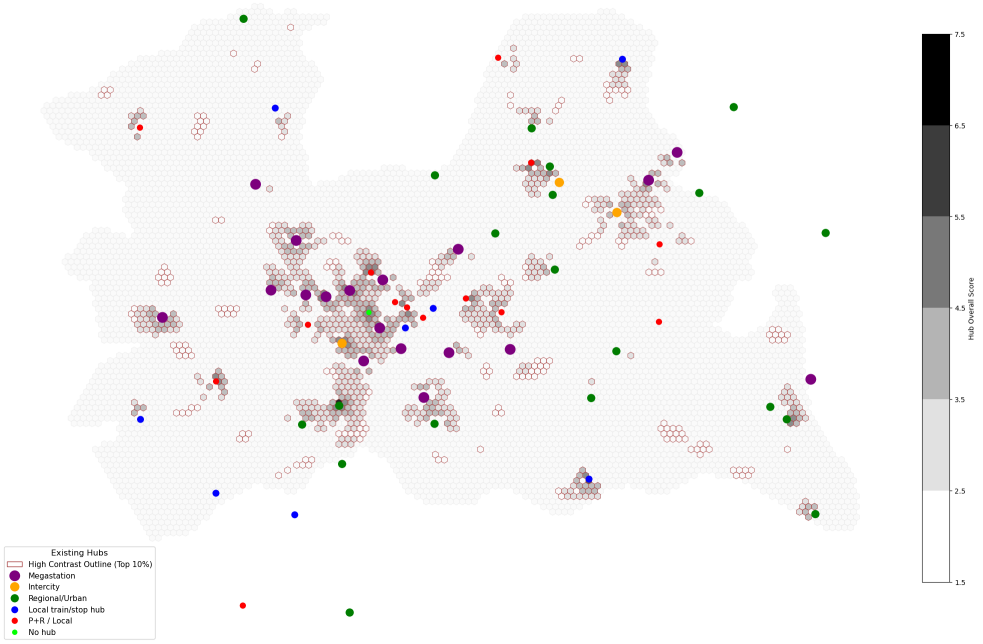


1. **Mature zones with high infrastructure coverage** cluster around central Utrecht. These areas score high on PT access, hub quality, and OV-fiets integration, suggesting a stable supply network with little room for expansion.
2. **Transition zones with growing housing demand** show mid-range hub scores and partial PT integration. These areas may not yet be underserved, but require **anticipatory mobility planning** to keep pace with urban development.
3. **Peripheral areas with low hub presence** emerge clearly through the absence of OV-fiets access and low `hub_overall_score`, despite having notable projected housing or job growth. These are **high-opportunity zones for hub creation or upgrades**.

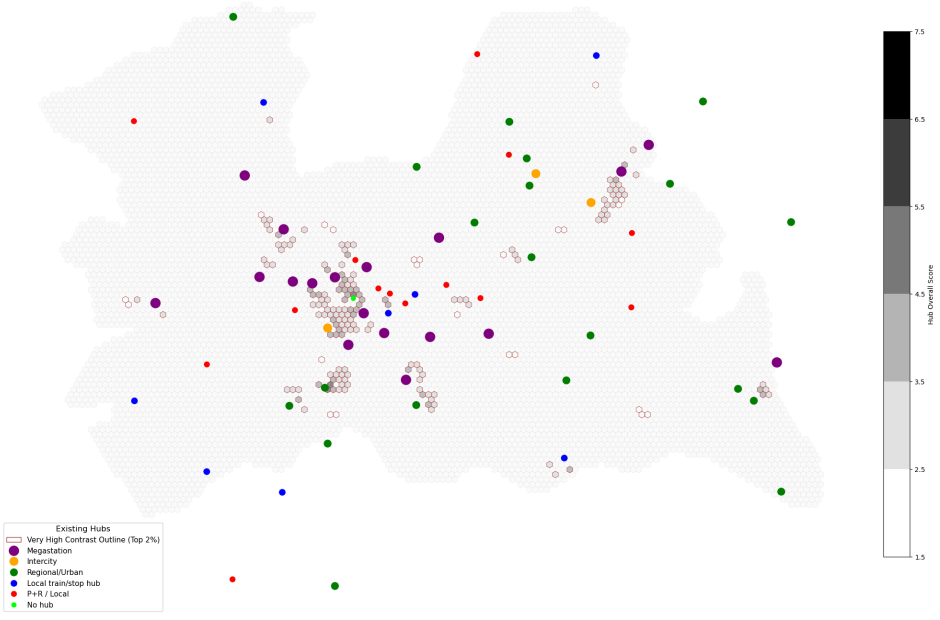
Gap Score and Prioritization

To operationalize these findings, an **infrastructure gap score** (`gap_score_infra`) was computed by combining **demand and supply indicators**. It ranks each hexagon by the extent to which **urban growth outpaces infrastructure availability**. A selection was then made based on the computed scores to identify high-priority areas for intervention:

Top 10% Demand–Supply Gap Areas
(Priority Candidates for Mobility Hub Improvement)



Top 2% Very High Gap Areas
(Critical Zones for Infrastructure Intervention)



Key Insights: Gap Area Distribution by Hub Score

Hub Score	Hexagons (Total Gap Areas)	In Top 2%	In Top 10%

3	426	112	426
4	233	41	233
2	199	23	199
5	34	3	34
6	1	0	1
7	1	0	1

1. **Most underserved hexagons fall in the 2–4 hub score range**, indicating areas that have basic infrastructure but are not keeping up with projected demand.
2. **Score 3 dominates both Top 2% and 10% gaps**, suggesting these zones are within reach of mid-tier hubs but lack supporting modes like shared bikes or vehicles.
3. **Score 2 areas are on the edge of networks**, pointing to early-stage development zones where access remains limited.
4. **Even Score 4 zones appear frequently**, signaling that existing hubs may be overloaded or not well connected to local development patterns.
5. **High-score zones (5–7)** are rarely flagged, confirming that truly multimodal areas are generally not underserved.

→ Focus planning efforts on zones scoring 2–4: areas with partial infrastructure that need expansion or multimodal integration to support growth.

▼ Conclusion and Strategic Recommendations

This study applied UMAP + HDBSCAN clustering across three curated feature sets to explore mobility demand, infrastructure gaps, and spatial mismatches in the Utrecht region. Each feature set served a different planning lens and helped reveal where shared mobility infrastructure is most needed.

Cross-Cutting Insights

1. Urban Growth Is Outpacing Infrastructure

Across all feature sets, clusters with high housing projections and growing job concentrations often lacked equivalent public transport access or proximity to shared mobility hubs. This is especially pronounced in the **"Housing and Jobs – Urban Demand"** and **"Infrastructure Availability"** feature sets.

→ *Strategic action*: Prioritize hub creation and PT integration in high-growth peripheral zones.

2. Fragmentation Indicates Micro-Level Gaps

In the **"First and Last Mile Need Zones"** feature set, clusters were highly fragmented, indicating that last-mile challenges are not just regional but highly localized. These clusters highlight the importance of granular planning.

→ *Strategic action*: Deploy flexible and modular shared mobility services in fragmented zones, focusing on short-distance connections.

3. Stable Core vs. Transitional Fringe

The central urban core appears consistently well-served, with limited opportunity for expansion. However, clusters on the fringe — especially where demand is rising — show recurring infrastructure lag.

→ *Strategic action*: Develop a forward-looking network plan for **transition zones** to prevent future mismatches.

Priority Zones for Mobility Investment

We operationalized the gap analysis in two ways:

- **Top 10% Demand–Supply Gap Areas**: Strong candidates for targeted upgrades in existing infrastructure and hub capacity.
- **Top 2% Critical Zones**: Urgent cases where current infrastructure is least aligned with expected demand — a shortlist for pilot projects and funding prioritization.

Together, these insights form a data-driven foundation for **strategic investment in Utrecht's transport infrastructure**, supporting a more equitable, multimodal, and future-ready mobility system.