Analyzing commuting patterns using mobile phone data: the case of Cluj-Napoca's rapid economic transformation

A Methodological Approach for Home-Work OD Matrices

Vlad I. Alexe, Norbert Petrovici, Camelia F. Andor
Interdisciplinary Centre for Data Science ,Babeș-Bolyai University

BABEȘ-BOLYAI UNIVERSITY

the interdisciplinary
centre for data science
—

Study Purpose & Approach

Analyze **commuting patterns** in Cluj County Romania to support urban planning and infrastructure development, addressing the mobility demands arising from the region's rapid growth.

Develop and validate a **methodological approach** to generate commuting OD matrices using mobile phone data.

A **network analysis** framework is applied to examine both structural and temporal stability in commuting flows

Cluj - Napoca

Second-Largest City in Romania Population: 320,000 (with 130000 in the metropolitan area)

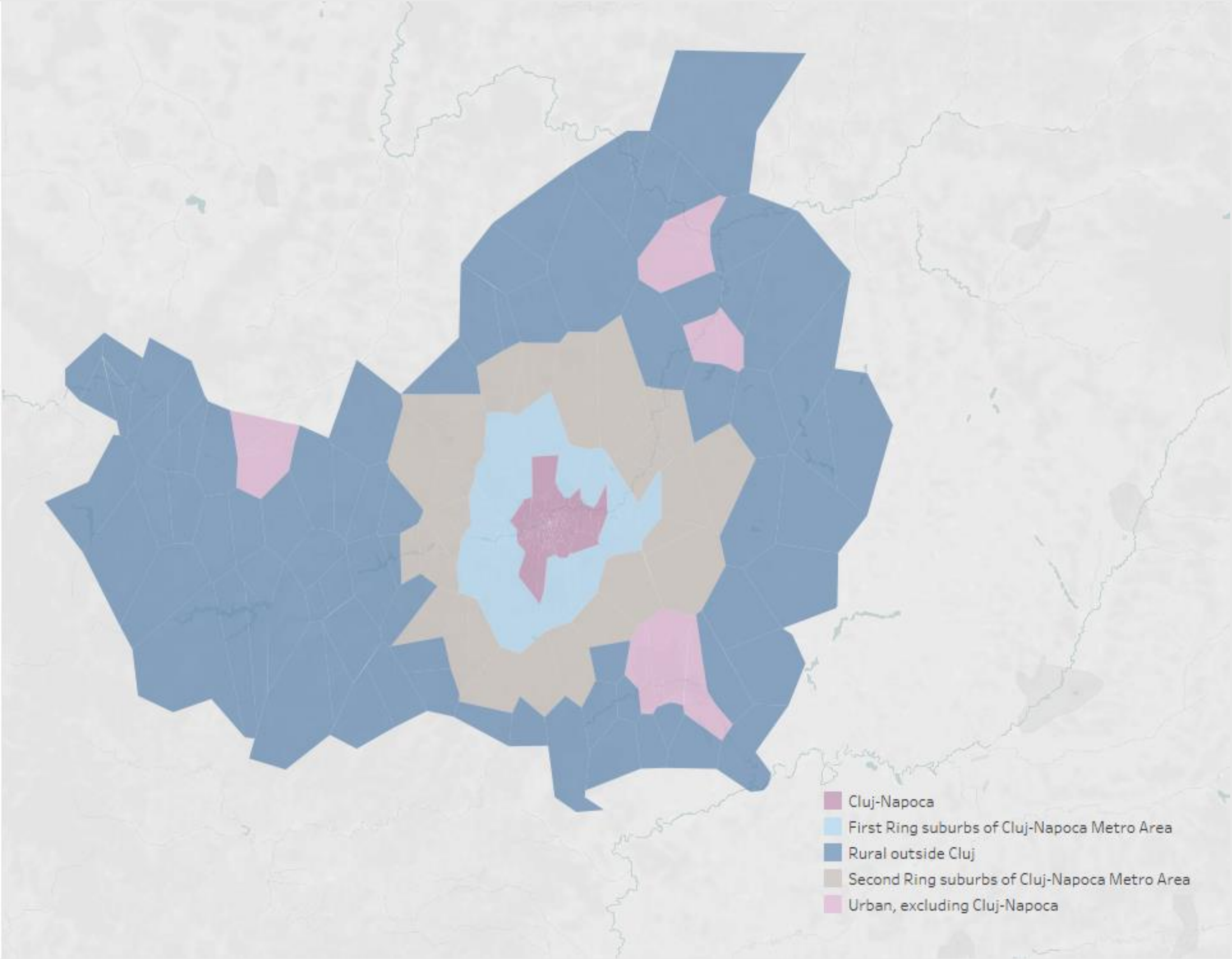
Economic Growth:
- Substantial increase in GDP by 4.5 times over the past 20 years
- Fastest-growing economy in the EU

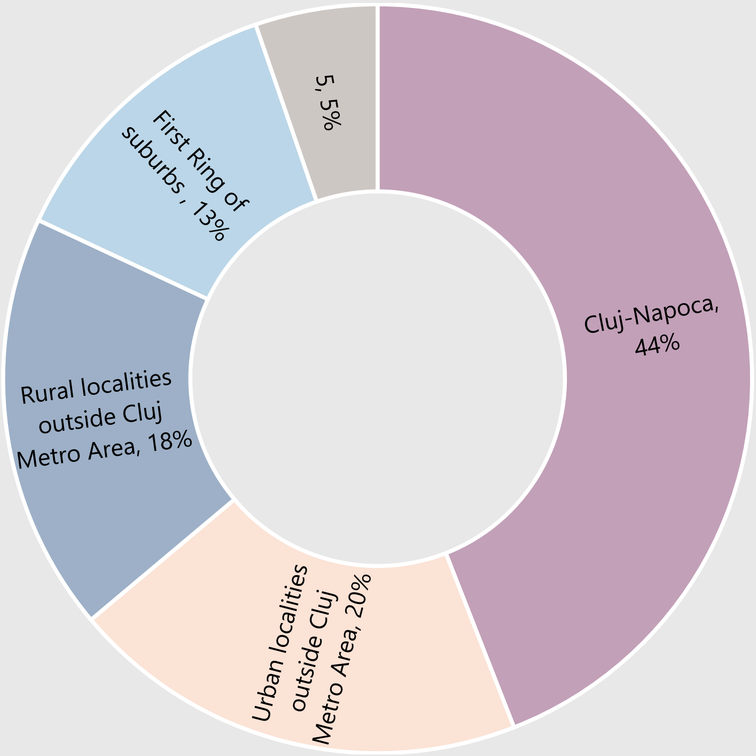
Urban Development:
- Metropolitan area population increased by 46% in the last two decades
- Notable growth in Florești (fivefold) and Jucu (50%)

Real Estate Boom:
- 90% rise in real estate prices in the past 7 years

Socio-Economic Dynamics:
- **Average Monthly Income:** €1,445 (18% above national average)
- **Employment:** 189,000 employees
- **Aggregated Turnover:** €14,263 million (second only to Bucharest)

Cluj County





Population:

Cluj –Napoca : 327212

Urban localities outside Cluj Metro Area : 147055

Rural localities outside Cluj Metro Area: 134394

Cluj- Napoca Metro Area : 133929

Florești, Cluj

- Largest rural locality in Romania
- Population ~ 50.000
- 7x growth in the last 20 years
- Chaotic urbanism with poor infrastructure and access


From Florești to Cluj -Napoca

Mobility Data - Network Signaling Data (NSD)

NSD logs user locations on events such as calls, SMS, handovers, attachment/detachment events, and internet connections

The dataset includes telecom data from Orange Romania, covering 36% of the national network. Data was collected from **339 cell tower sites**, with privacy ensured through anonymized user IDs and temporal data discretized into 15-minute intervals.

The raw dataset consists of **40.3 GB** of compressed data, spread across **13,027 Parquet files.**

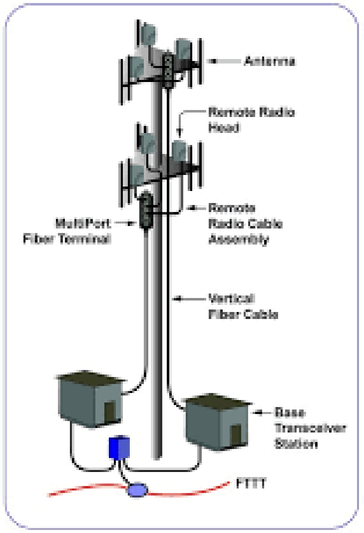
Secondary Data Integration & Spatial Analysis

Population estimates were derived from the Romanian National Statistics Institute and the Global Human Settlement Layer.
Employment data was provided by the Cluj Work Inspectorate, detailing locations and employee counts by sector for companies in Cluj.

Spatial representation was constructed by intersecting Voronoi cells with hexagons using Hexagonal Hierarchical Spatial Index (H3). Employment data was aggregated within Voronoi polygons, and hexagonal boundaries were aligned with official administrative limits to ensure accurate population weighting.

LIMITATIONS OF NSD DATA

- Tower Allocation Changes
- Ping-Pong Effects between antennas
- Non-Human Transactions
- Insufficient Stay Locations
- Inconsistent time frames
- Uneven antenna coverage
- Cell tower density difference between rural and urban
- Overlapping antenna coverage
- Processing power requirements



DATA COLLECTION AND PREPROCESSING PHASE

Step 1: Data Acquisition and mobility database

- *Data Source*: Network Signalling Data (NSD) from Orange Romania, Cluj County for one month
- *Data Characteristics*: 40.3 GB of compressed data, distributed across 13,027 Parquet files.
- Local MongoDB server for handling raw and preprocessed data nearly 1 billion records

Step 2: Data Preprocessing

Handling noise and adapt data to computational capabilities.

- *Mongo Pipeline* to to format the data into a structure consisting of user ID, site location, and timestamp.
- *Filtering*: Removal of transient data (e.g., quick pass-throughs) – based on event frequency distribution for every user. 60% of all data were of user with less than 35 events
- *Compression*: keeping only the first and last event at each consecutive site

Step 3: Spatial Delineation

Tools : ArcGIS, Python

- Voronoi Polygons: Used to delineate each cell tower's coverage area, which helps approximate antennas coverage and refine movement trajectories.
- Centroid Analysis: Centroids of Voronoi polygons serve as spatial units for further analysis.
- Distance matrix between all the centroids of the Voronoi polygons – geodesic distance and road distance

DATA ANALYSIS PHASE

Step 1: Home and Work Location Detection

Home Detection:

- *Nighttime Data Processing:* Analysis of events between 00:00 AM and 08:00 AM to identify potential home locations.
- *Clustering:* Hierarchical clustering using Ward's method to group location points and identify home locations based on cumulative time spent at each cluster location.

Work Detection:

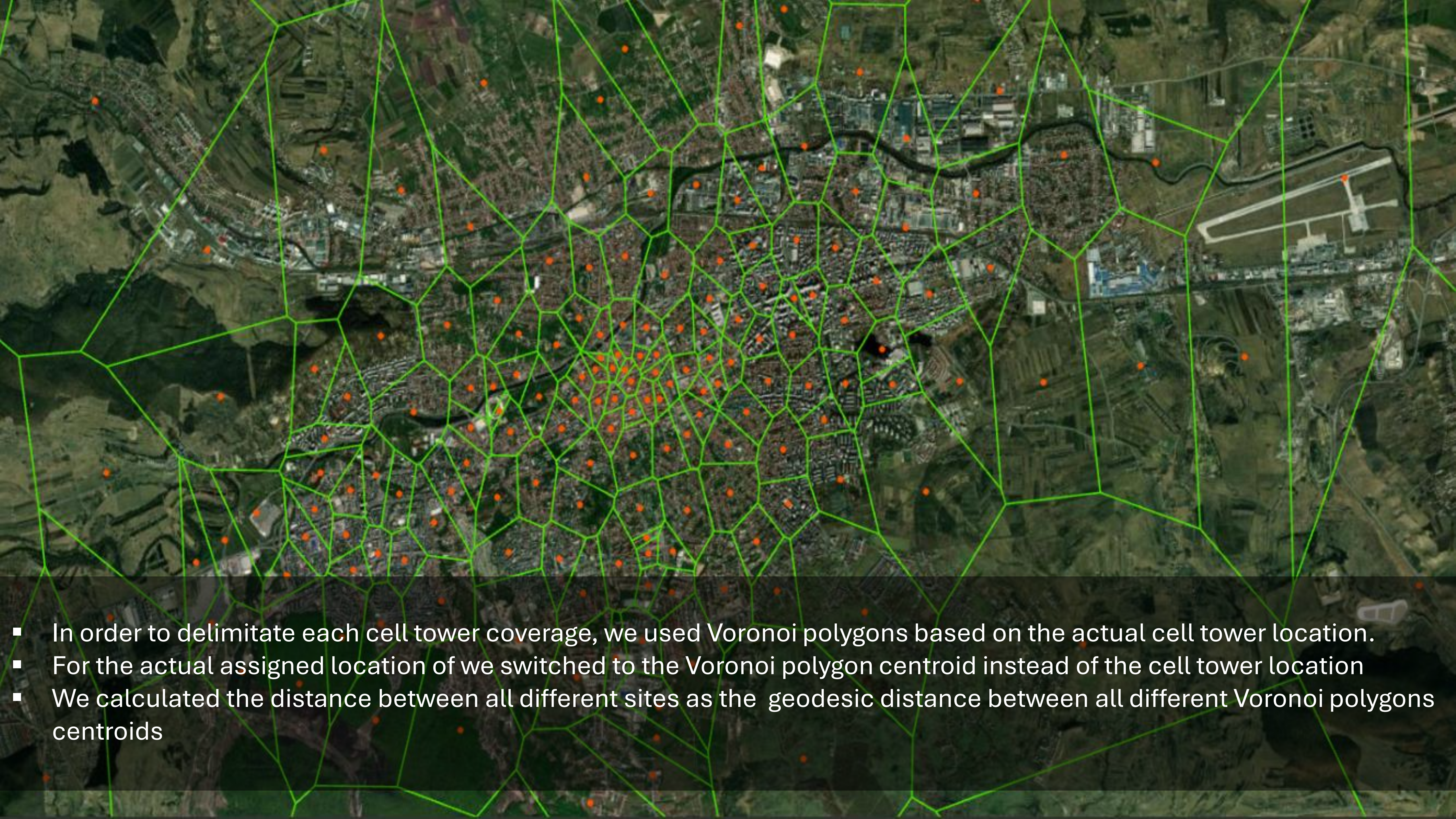
Daytime Data Processing: Analysis of events between 08:00 AM and 08:00 PM to identify work locations.

Clustering and Analysis: Similar clustering method used for detecting work locations, focusing on stationarity during daytime hours.

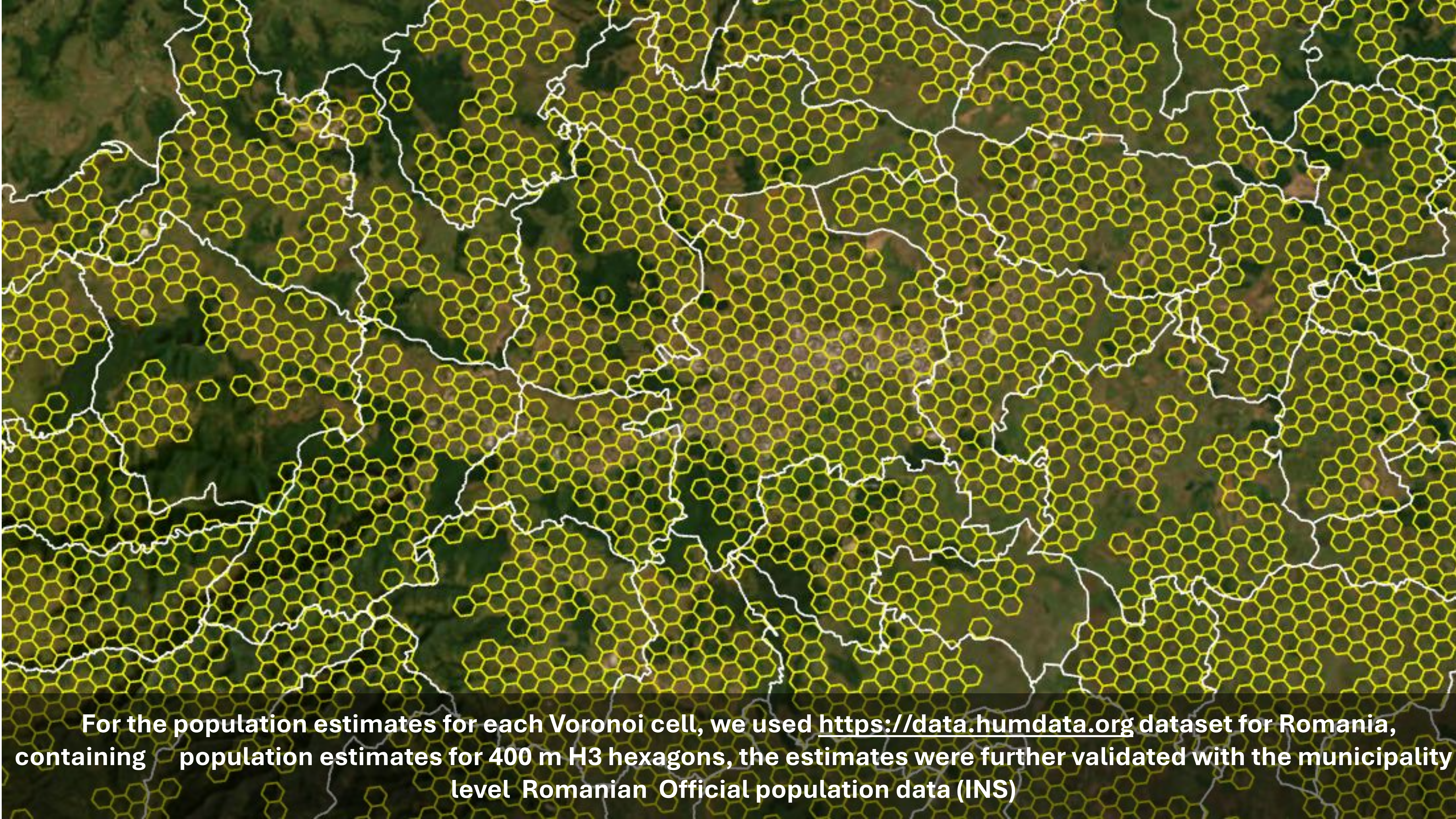
- **Step 2: Commuting network graph**
- *Origin-Destination Matrix* Created between the last time at home after 04:00 and first time at work location.
- Separate OD based on weeks, weekdays and weekends
- *Commuting Graph from OD matrix:* An undirected graph for every segment of the OD and a general graph for weekends and weekdays
- Added *node attributes* (locality type, population and employees count) and edge attributes (commuting counts, average speed, average distance)

Step 3: Analysis

- *Home and work location validation* using cross – variogram – accounting for spatial dependency
- *Community detection* using Leiden algorithm – added as node atributes
- *Community stability analysis* across graphs using ARI
- *Network metrics* – degree, betweenness, closeness – added as node attributes
- *Visualizations* – maps for communities, degree centrality, home and work densities



- In order to delimitate each cell tower coverage, we used Voronoi polygons based on the actual cell tower location.
- For the actual assigned location of we switched to the Voronoi polygon centroid instead of the cell tower location
- We calculated the distance between all different sites as the geodesic distance between all different Voronoi polygons centroids



For the population estimates for each Voronoi cell, we used https://data.humdata.org dataset for Romania, containing population estimates for 400 m H3 hexagons, the estimates were further validated with the municipality level Romanian Official population data (INS)

MobTrack - Key Modules and Features

Python library for mobility analysis: https://github.com/CIDS-UBB/MobTrack (the library is not public yet)

Preprocessing:

- Base Dataframe Creation: Foundation for simplified, efficient data analysis.
- Data Preprocessing: Functions for various preprocessing needs.
- Filters: Remove noise and limit the data scope as needed.

Geo Module:

- Proximity Clustering: Groups nearby sites to enhance spatial accuracy.
- Distance Calculation: Measures travel between clusters.
- Tessellation: Assign locations using grid, hexagons, or Voronoi methods.

Clustering:

- Home Sites: Detects residential location, Work Sites: Identifies work locations for commuting.
- Incremental Clustering, DBScan, Hierarchical Clustering

MongoDB Integration: Mongo Utils: Efficient data upload, storage, and retrieval.

Statistics: Descriptive Analysis: Provides metrics to understand trends and distributions.

Models: Machine Learning and Statistical Models: Applied to mobility analysis.

Network: Functions for network construction and analysis

MobTrack

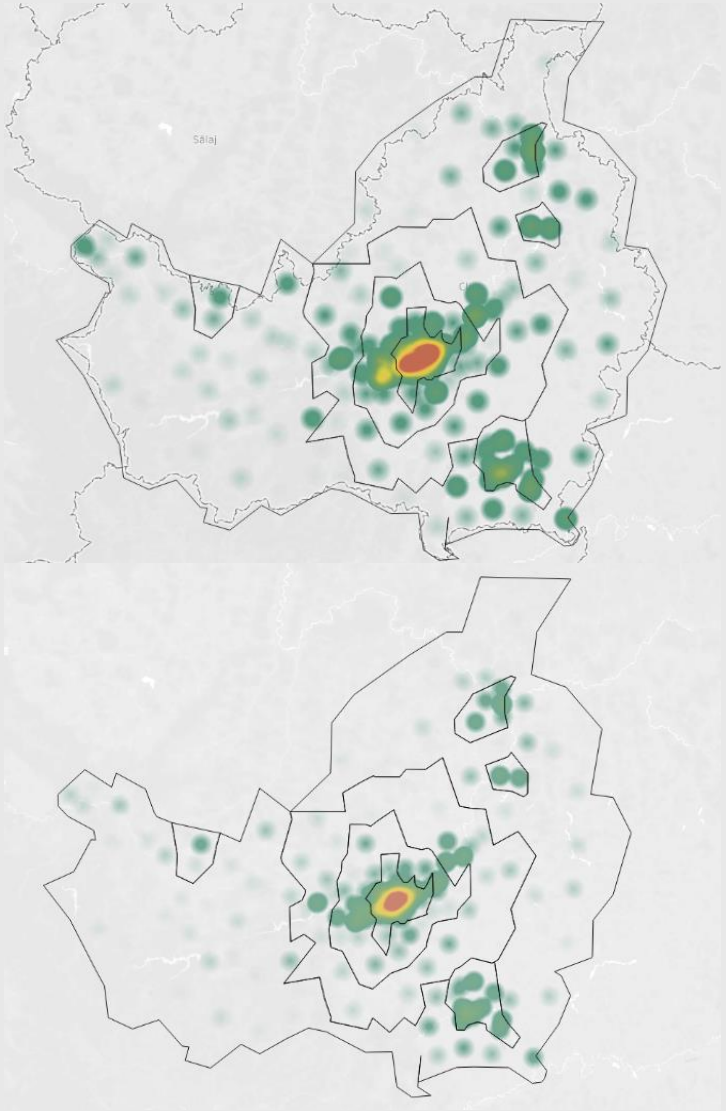
Detected home locations (up) and work locations (down)

Home Locations: High density in urban areas, especially the central urban core in Cluj–Napoca city and First Ring Metro Area, lesser densities in other urban localities, with significantly lower densities in rural zones.

Work Locations: More concentrated in the center of Cluj-Napoca and the other urban localities.

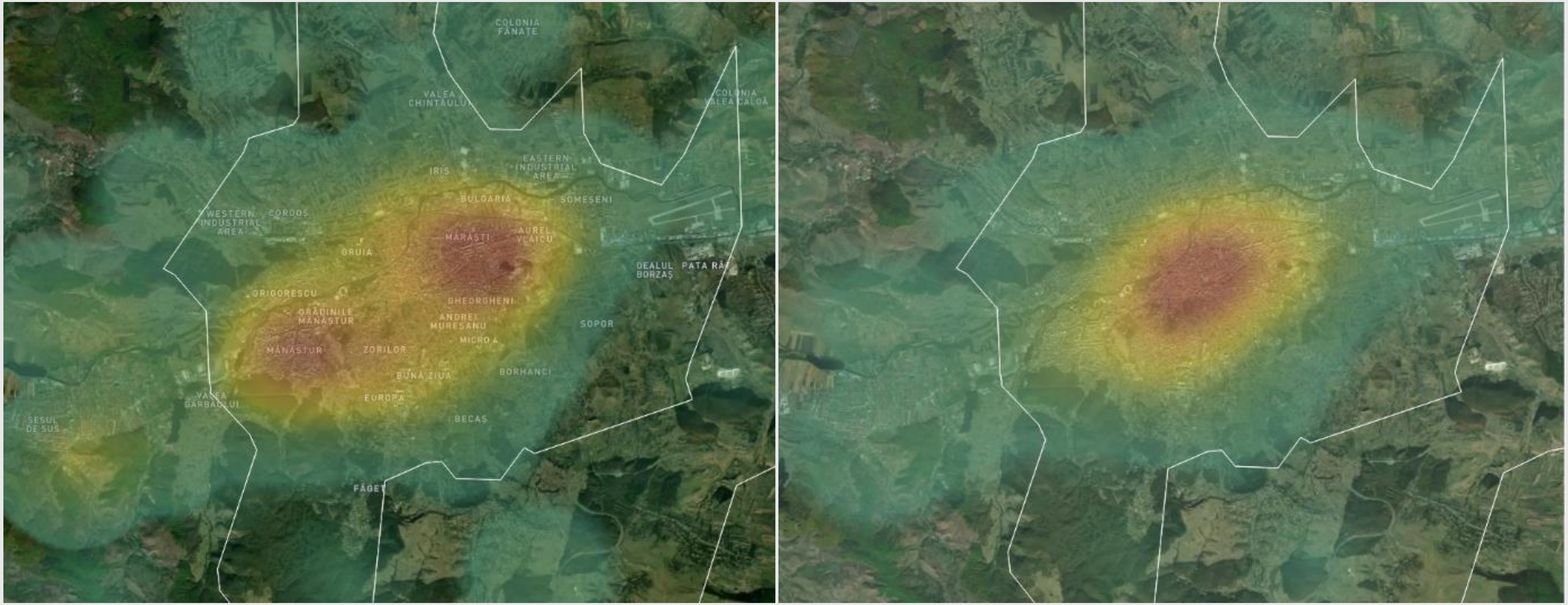
Key Insight: Differences in clustering reflect commuting patterns and center – periphery dynamics.

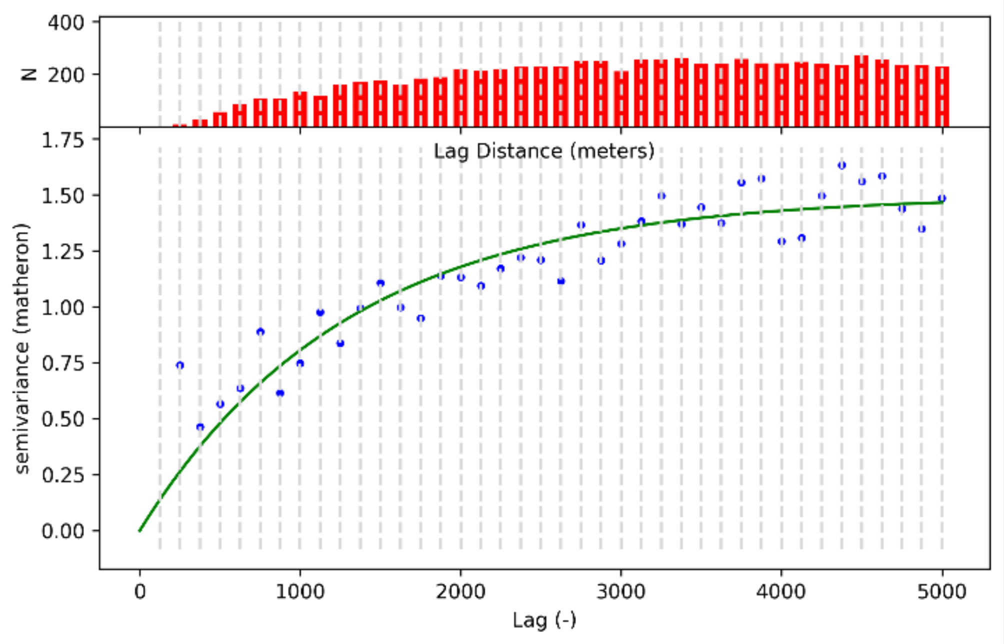
Validation: Patterns align with known population and workplace distributions, supporting accuracy of home/work detection.

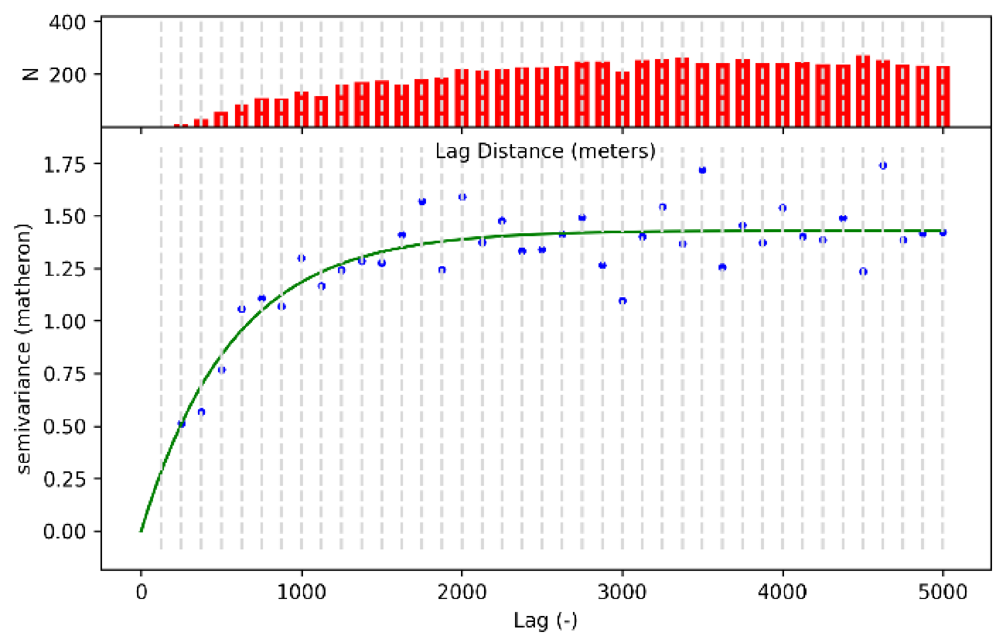


Home location density, work location density for Cluj-Napoca city

The left map shows residential clusters in Mănăștur and Mărăști, while the right map, workforce concentration in the city center. These patterns align with known population and employment distributions, validating the detection methodology.







Validation Approach Using Cross-Variograms:

Method: Cross-variograms with scikit-gstat assessed spatial dependency between detected home/work locations and population/employee counts by calculating semi-variance using an exponential model.

Model Fit: The exponential model stabilizes at a sill, suggesting reduced correlation at greater distances, validating the clustering accuracy.

Spatial Dependency: Steep semi-variance increase up to ~2000 meters indicates strong spatial dependency between detected locations and actual densities.

Accuracy Confirmation: Strong spatial correlation at short distances confirms accurate clustering of residential and work locations.

BABEȘ-BOLYAI UNIVERSITY

the interdisciplinary centre for data science

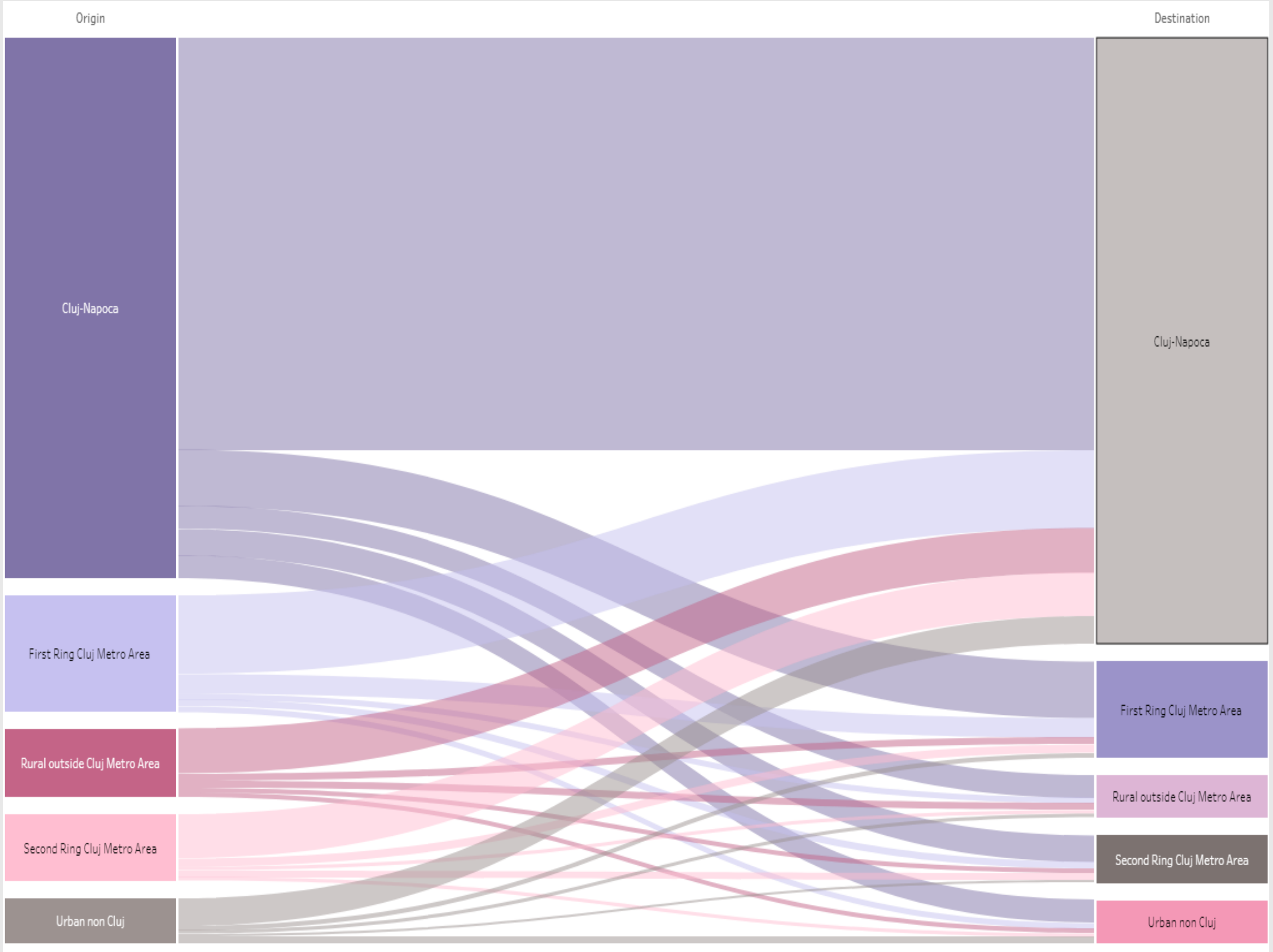
Home – work route

- Adjust site changes based of home and work proximities
- Find the time the user left home during the morning
- Find the time the user first arrived at work
- Compute time to work and distance to work
- Generate home-work OD matrix

BABEȘ-BOLYAI UNIVERSITY

the interdisciplinary centre for data science

OD matrix –commuting flows

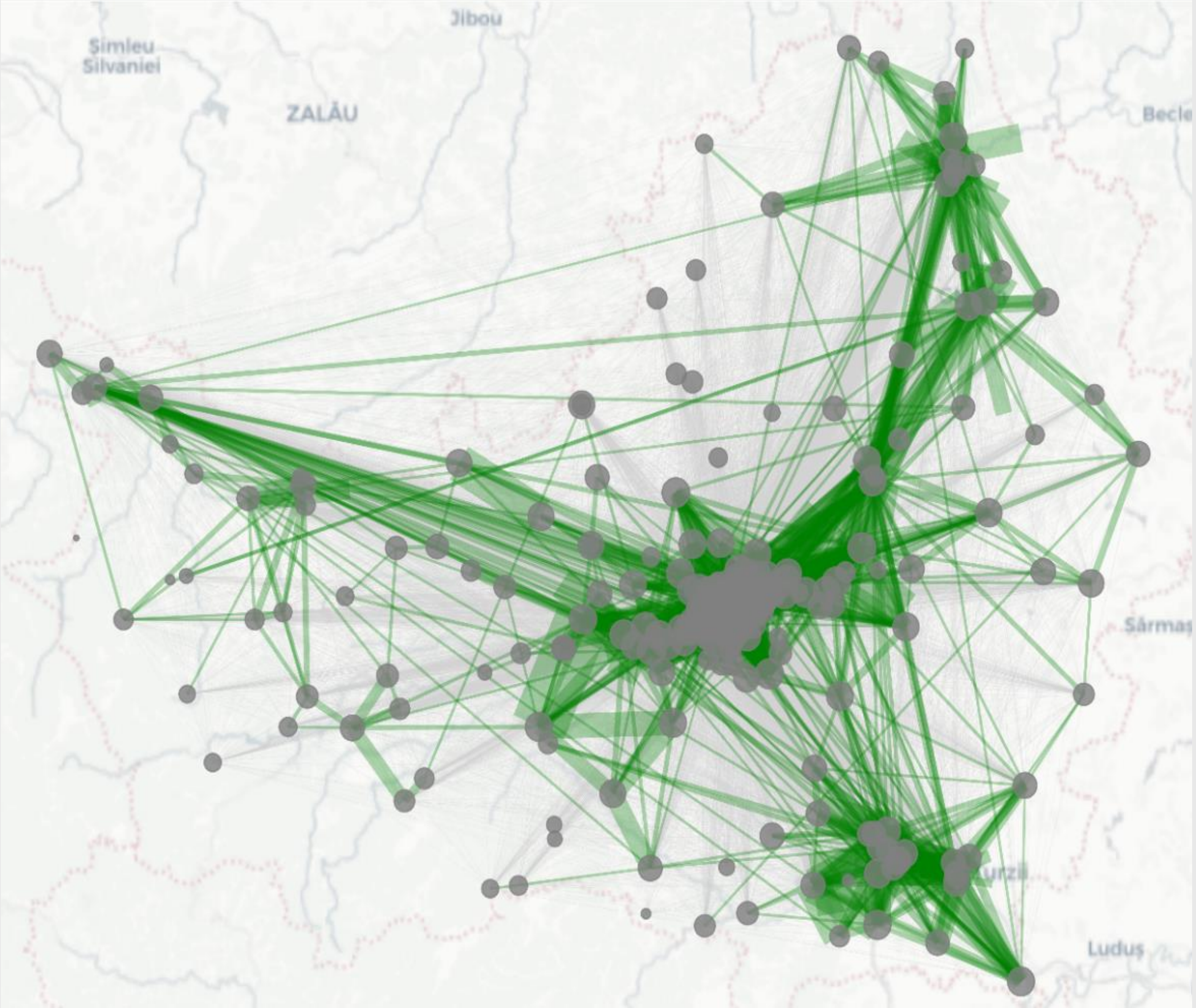


Dominance of Cluj-Napoca:
Cluj-Napoca serves as the primary origin and destination for commuting flows, with the largest share of connections both starting and ending within the city.
Strong Urban-Suburban Linkages: Significant flows connect Cluj-Napoca to the First Ring Metro Area
Weaker Rural Integration: Flows to and from rural areas (outside the metro area) are less prominent
Regional Dynamics: Second Ring Metro Area and other urban non-Cluj localities show moderate commuting connections,

Commuting network

The graph of Voronoi centroids linked by commuting flows highlights Cluj-Napoca as a major employment hub, along with other smaller hubs in urban localities.

There are significant commuting flows between Metropolitan Area localities and Cluj-Napoca, as well as along major roads connecting urban localities within the county.



bubble size: node degree; edge size: average number of commuters

BABEȘ-BOLYAI UNIVERSITY

the interdisciplinary centre for data science

Communities Around Cluj-Napoca:

Five well-defined geographic clusters detected using the Leiden algorithm.

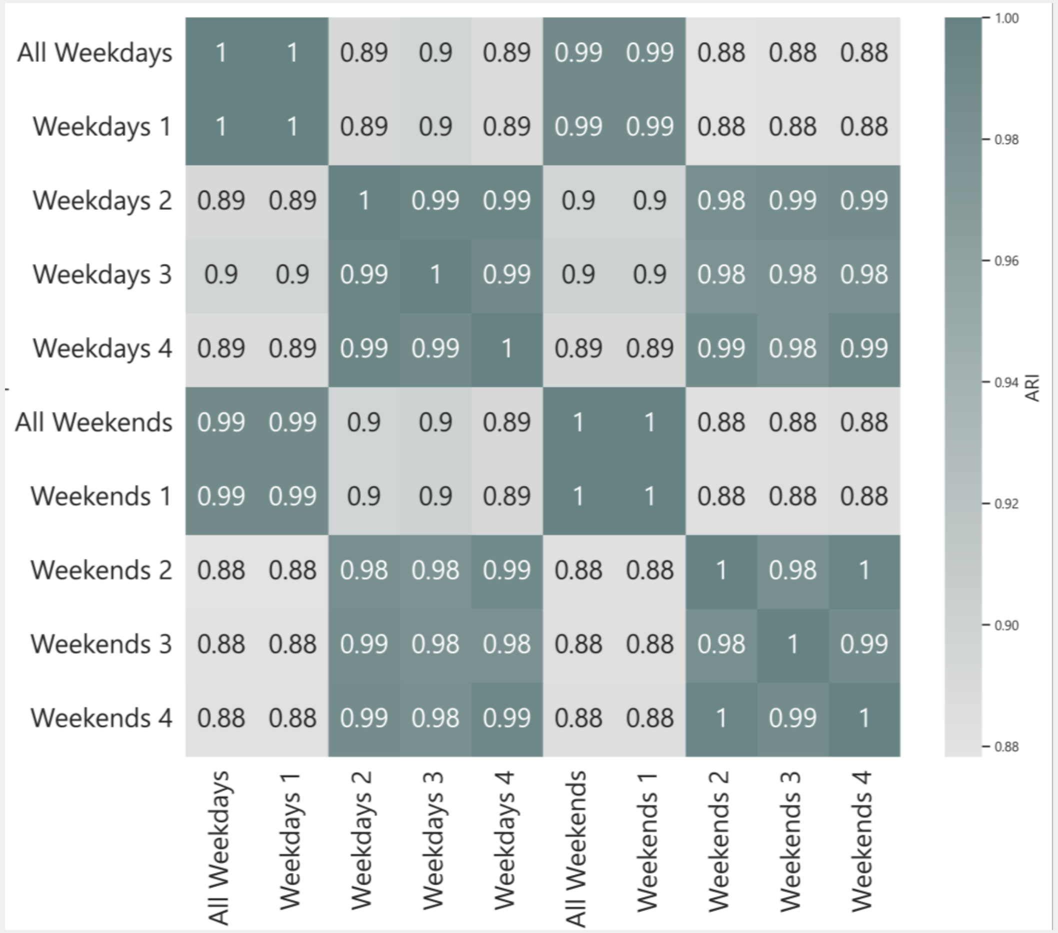
Cluj-Napoca as Central Convergence Point: The central hub where multiple communities connect.
Geographic and Infrastructural Influences: Communities separated by cardinal directions (south, northwest, north, northeast), shaped by traffic routes and natural features like hills and mountains.

Key Insights for Connectivity Improvements:

Distinct Community Identities: Communities maintain unique identities due to geographic and infrastructural barriers.
Implications for Planning: Highlighting the need for targeted infrastructure investments to enhance inter-community connectivity, especially where natural barriers limit access.

BABEȘ-BOLYAI UNIVERSITY

the interdisciplinary centre for data science



Adjusted Rand Index (ARI): Heatmap

Community Stability analysis

Stability Assessment: ARI used to measure consistency in community structures across weekdays, weekends, and individual weeks.

Temporal Consistency:

- **High ARI (~1)** within weekdays and weekends indicates stable clustering.
- **Moderate ARI (~0.88)** between weekdays and weekends reflects shifts due to differing activities and travel behaviors.

Event-Specific Dynamics:

- **Lower ARI for "Weekdays 1"** corresponds with the start of the academic year, indicating temporary shifts due to student arrivals.

Insights:

- **Overall Stability**: Community structures are mostly stable, with minor shifts during specific events like the academic year start.
- **Validation**: High stability supports the accuracy of the OD matrix construction for temporal analysis.

Conclusion and Key Insights

Cluj-Napoca as a Commuting Hub: Confirmed as the central node for regional commuting, characterized by high degree centrality, with strong center-periphery connectivity and limited polycentricity.

Community Analysis: Identified five distinct commuting clusters; Cluj-Napoca and surrounding urban areas are densely connected, while rural zones are more isolated, indicating a need for targeted infrastructure development.

Methodological Contributions: Developed MobTrack for custom mobility data processing and analysis, validated through cross-variograms and ARI-based stability analysis

Policy Implications: Strengthen transit links by focusing on rail, tram, and high-frequency bus services between Cluj-Napoca and nearby municipalities, along with developing a metropolitan beltway to improve connectivity.

Study limitations and future directions

Limitations:

Data Coverage: 36% of network data—limited representativeness.
Temporal Scope: Single month data—lacks seasonal insights.
Home detection validation: The focus on only on one home detection algorithm.
Voronoi tesselation: Voronoi polygons do not accurately represent cell tower coverage.
Data Filtering: Important mobility events may be excluded.

Future Directions:

Expand Temporal Data: Include long-term data for better seasonal trends.
Broader Integration: Add public transit, mobility census data, employees by economic sector, other points of interest data
Enhance MobTrack: Improve clustering algorithms and adapt for multiple data sources.
Increase Accuracy: Use cell tower coverage maps data to increase location accuracy, experiment with 1 km grid cells instead of Voronoi polygons
Broader Application: Validate approach in other metropolitan regions

Selected references

- Alexander, L., Jiang, S., Murga, M., & González, M. C. (2015). Origin-destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies*, *58*, 240–250. https://doi.org/10.1016/j.trc.2015.02.018
- Calabrese, F., Di Lorenzo, G., Liu, L., & Ratti, C. (2011). Estimating origin-destination flows using mobile phone location data. *IEEE Pervasive Computing*, *10*(4), 36–44. https://doi.org/10.1109/MPRV.2011.41
- Fekih, M., Bellemans, T., Smoreda, Z., Bonnel, P., Furno, A., & Galland, S. (2021). A data-driven approach for origin--destination matrix construction from cellular network signalling data: a case study of Lyon region (France). *Transportation (Amst), 48*(4), 1671–1702. https://doi.org/10.1007/s11116-020-10108-w
- H3. (2023). *Hexagonal Hierarchical Geospatial Indexing System*. https://h3geo.org/.
- Papuc, G., & Vizireanu, N. (2023b). Types of Mobile Telecommunications User Location Errors Encountered in Urban Mobility Statistical Analyses. *2023 15th International Conference on Electronics, Computers and Artificial Intelligence (ECAI), 1–4.* https://doi.org/10.1109/ECAI58194.2023.10193947
- Vanhoof, M., Reis, F., Ploetz, T., & Smoreda, Z. (2018). Assessing the Quality of Home Detection from Mobile Phone Data for Official Statistics. *Journal of Official Statistics*, *34*(4), 935–960. https://doi.org/10.2478/jos-2018-0046

Thank You!

Feel free to reach out for questions or further discussion:

vlad.alexe@ubbcluj.com