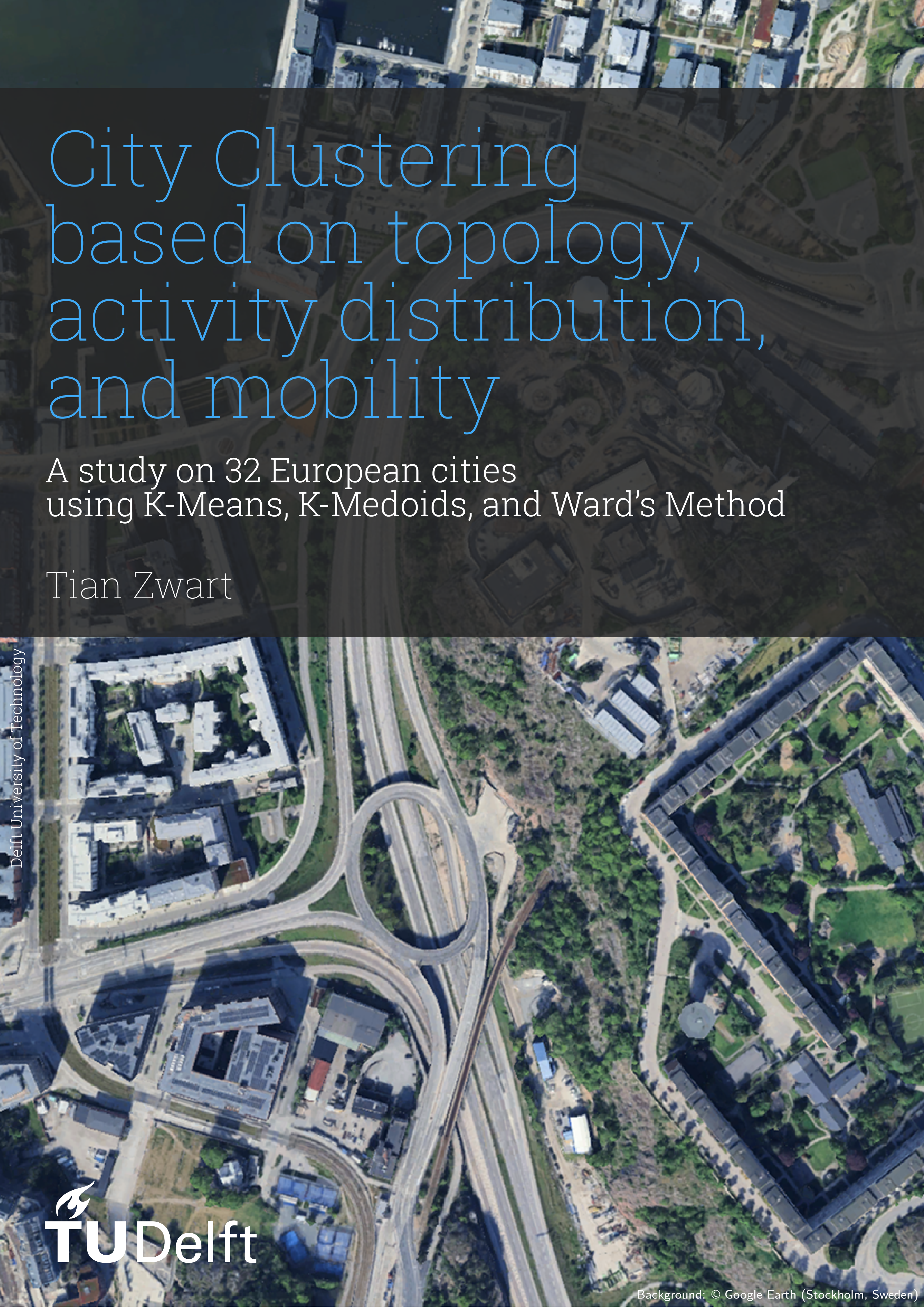
City Clustering based on topology, activity distribution, and mobility

A study on 32 European cities using K-Means, K-Medoids, and Ward's Method

Tian Zwart



Background: © Google Earth (Stockholm, Sweden)

Delft University of Technology

City Clustering based on topology, activity distribution, and mobility

A study on 32 European cities using K-Means, K-Medoids, and Ward's Method

by

Tian Zwart

to obtain the degree of Master of Science
in Civil Engineering - Traffic and Transport Engineering,

at the Delft University of Technology,

to be defended publicly on Wednesday, May 28, 2025 at 12:45 PM.

Student number:	4955609	
Project duration:	October, 2024 – May, 2025	
Thesis committee:	Dr. M. Snelder,	TU Delft (Chair)
	Dr. ir. I. Martínez,	TU Delft (Daily Supervisor)
	Prof. L. Leclerq,	TU Delft

TU Delft

Preface

This report marks the final chapter of my time as a student at TU Delft. From starting the Civil Engineering bachelor's in 2018 to concluding my studies in 2025, the past seven years have been a period of growth, exploration, and challenge, both academically and personally. I was drawn to Civil Engineering by an interest in the infrastructure that shapes our environment, and this interest evolved into a focus on Traffic and Transport Engineering, a field that directly connects this infrastructure with everyday human behavior.

I chose this research project for its holistic perspective on urban mobility. I hope its practical insights can support cross-city learning and contribute to improving urban life. While research can sometimes feel abstract, I believe its relevance becomes clear when tied to real-world challenges. I have always been drawn to understanding the bigger picture, how different elements interact and shape one another. Though there is always room to improve, I am proud of the insights this work offers to the analysis and comparison of urban transport systems.

Of course, I could not have done this alone. First and foremost, I want to thank my supervisors. A special thanks to dr. ir. I. Martínez for her patience, especially during the early phase when I had to take some time off. You encouraged me to determine my own direction in this research, which, combined with your advice, helped me grow and learn independently. Our meetings always helped me refocus and move forward. I also want to thank dr. M. Snelder for chairing my graduation committee, your guidance and feedback were very helpful, and you ensured the process ran smoothly. Finally, I thank prof. L. Leclercq for joining the committee and contributing valuable feedback. It was much appreciated that you were in Delft for the key meetings, great timing!

I am also very thankful to my family, my boyfriend, the friends who supported me throughout my studies, both those I met in Delft and those from outside, and everyone who was pushing through their thesis alongside me in the thesis room. This year began with some difficult moments, but thanks to your support, I am able to end it on a high note.

To all readers, I hope this report offers both insight and inspiration. Let's remain open-minded, not only in research, but also in how we learn from and connect with one another. That mindset is essential to building a better environment for us all.

Tian Zwart
Delft, May 2025

Summary

Cities face growing challenges driven by rapid urbanization and climate change, while striving to maintain or improve the quality of life in the urban environment. Transportation plays a central role in these challenges, occupying valuable space and contributing to harmful emissions, especially in car-oriented cities. Although a wide range of solutions exists, their effectiveness strongly depends on the urban context. Poorly adapted strategies can lead to unintended consequences. To address shared challenges more efficiently, cities can benefit from learning from similar urban contexts rather than reinventing the wheel themselves. However, identifying comparable cities remains a difficult task.

Previous research has classified cities based on individual domains such as road geometry, network topology, land use composition, or mobility patterns. While these studies provide valuable insights, they typically capture only a single aspect of urban transport systems. As a result, the influence of other factors and the interdependencies between domains are not considered, limiting the practical applicability of the findings. Recognizing these interconnections can reveal important patterns and offer valuable lessons for urban development.

This research addresses that gap by developing a clustering framework that integrates multiple domains, combining structural characteristics, activity distribution, and mobility behavior into a holistic analysis. The societal aim is to support more targeted cross-city learning and strategy development by identifying groups of cities that share underlying transport and urban structure features. The main research question for this research is formulated as:

How can the application of multiple clustering methods reveal distinct groups of European cities based on road network, activity distribution and mobility characteristics?

To answer this question, 32 European cities were selected based on consistent data availability. A wide range of indicators was collected and calculated to characterize each city, covering five domains: road topology, population distribution, economic activity distribution, mobility behavior, and congestion. After ensuring comparability through standardized boundaries and indicator definitions, the dataset was reduced through a correlation analysis and Principal Component Analysis (PCA), retaining essential variation while minimizing redundancy in the dataset. Clustering was then performed using three different methods: K-Means, K-Medoids, and Ward's Method. The results were evaluated using complementary metrics, including the silhouette score, the Adjusted Rand Index (ARI), and the Jaccard Similarity.

The results showed that clustering cities based on combined structural, functional, and behavioral indicators produced stable and meaningful groupings. Across all methods, the two- and seven-cluster solutions were the most consistent and informative. At the two-cluster level, a strong regional pattern emerged, separating Southern European cities from the rest. The seven-cluster solution showed complete agreement between methods, revealing distinct urban typologies characterized by differences in road structure, activity distribution, and mobility patterns. An overview of these clusters is provided in the table. In addition, the correlation analysis revealed several meaningful relationships between urban form, travel behavior, and congestion, highlighting the interdependencies that shape urban transport systems across European cities.

By integrating multiple domains and applying a systematic, stepwise methodology, this research contributes to the field of urban transport analysis in several ways. It fills an important gap by offering a more holistic classification of cities, demonstrating that meaningful urban typologies can be identified using publicly available data and multiple clustering methods. The combination of indicators covering multiple domains, transparent dimensionality reduction, the application of multiple clustering algorithms, and comprehensive result evaluation provides a replicable framework that can support future comparative studies and strengthen cross-city learning.

Summary of the seven clusters based on road network, activity distribution, and mobility characteristics.

Cluster	**Short Description**
Centralized Car-Oriented	Strong density contrasts, centralized, car-dominated with low congestion
Homogeneous Car-Oriented	Compact and uniformly dense, car-dominated with low congestion
Dense Multimodal	Extremely dense, multimodal, low car ownership and moderate congestion
PT-Oriented Congested	High public transport use, but also high congestion and car ownership
Well-Connected Bottlenecked	Well-connected but bottleneck-prone, moderate density and mixed mobility
Low-Density Concentrated	Dispersed residential pattern, centralized economic hubs, low PT use
Balanced Multimodal	Mixed densities, strong multimodal mobility, but notable traffic pressure

Building on the findings and limitations of this study, several recommendations for future research are proposed. Expanding the dataset to include historical city profiles would enable the analysis of urban development over time, enhancing the ability of cities to learn from transformations of other cities. Improving the quality and coverage of the input data, particularly for facility locations and mobility characteristics, would further strengthen the reliability and granularity of the clustering results. Incorporating explicit measures of polycentricity could provide deeper insights into how urban form influences travel behavior, while a more detailed exploration of the relationships between specific indicators could reveal underlying dynamics that shape urban transport systems. Applying this methodology to cities in other regions of the world could further test its relevance and uncover broader patterns of urban contexts.

In conclusion, this research demonstrates that clustering cities based on characteristics spanning multiple domains offers a powerful tool for understanding urban form and transportation systems. By improving cooperation, cross-city learning, and evidence-based planning, this approach can contribute to the development of more sustainable and resilient urban transport systems across Europe.

Contents

1

Introduction

Cities around the world are facing increasingly complex challenges. Urbanization is expected to rise from 56% in 2023 to 70% by 2050, which not only represents a relative increase: the total world population is also growing (World Bank, 2023). As urban populations grow, governments confront a wide array of issues, including housing shortages, energy demands, social inequality, and limited urban space for future development (United Nations, 2020; World Economic Forum, 2018). One system that relates to these challenges is urban transportation, essential not only for the movement of goods but mainly for supporting daily mobility. Whether commuting, visiting family or friends, doing groceries, engaging in leisure activities, transport systems shape how people experience and navigate urban life. Because transportation is deeply embedded in the physical and functional structure of cities, it provides a valuable perspective to analyze urban areas.

As more people move into cities, existing infrastructure, originally designed for smaller populations, is increasingly put under pressure by rising traffic volumes. This places additional stress on urban transportation systems, which are already facing spatial constraints. In many cities, between 30% and 60% of urban land is allocated to transportation infrastructure (Rodrigue, J., 2013), leaving limited room for expansion, especially as other urban functions, such as housing, public spaces and facilities, also compete for space. The majority of this infrastructure is dedicated to the road network, which primarily accommodates cars. However, roads serve multiple purposes: they are also used by buses and other public transport modes, and often include sidewalks for pedestrians and, in some cases, bicycle lanes. As cities face growing demands for mobility, while simultaneously seeking to reduce emissions and adapt to climate change, there is a pressing need to rethink how existing infrastructure can be transformed and contribute to a more sustainable and efficient transportation system within the constraints of limited urban space.

Adapting a city and its transportation network is an inherent part of urban development. Throughout history, streets have been modified in response to new mobility demands. When the tram was introduced, cities restructured parts of their networks to integrate it (Taplin, 1998), and the rise of the car led to widespread redesigns, followed later by further changes to improve pedestrian safety (Reid, 2022). However, changing infrastructure is a slow process, as cities are complex systems shaped by many interdependent factors (Ortman et al., 2020). This complexity results in cities developing in unique ways, creating significant variation in their spatial form and mobility systems.

Geographical context plays an important role in shaping urban transportation systems, as shown by Badhruudeen et al. (2022), who classified cities based on their geometric features. Road networks, the fundamental component of the transport system, can be described using topological indicators derived from graph theory. This approach represents the network as nodes (intersections) and edges (streets) (Wilson, 1996). Costa and Tokuda (2022) analyzed node-based indicators, including node degree metrics, the spatial distribution of nodes, and the variability in accessibility, to cluster 20 European cities. Similarly, Crucitti et al. (2006) introduced a range of centrality measures, including degree, closeness, betweenness, straightness, and information centrality, to identify important intersections and streets within the road network.

Yet the structure and functioning of urban transportation systems depend on more than just the road network. The distribution of population, along with the location of shops and workplaces, representing origins and destinations, significantly shapes travel patterns. These factors reflect the spatial structure of the city, such as whether it is organized around a single center or multiple sub-centers. In addition, the availability, diversity, and use of transport modes further influence how people move through the city, ultimately determining the performance and accessibility of the transportation system.

Despite the many differences in form, function, and mobility systems, cities also share key characteristics. With over 10,000 cities worldwide (Scruggs, 2020), it is inevitable that common patterns exist. As noted earlier, many urban areas face similar challenges, including rapid urbanization and the need to adapt to climate change. Given these shared challenges, in combination with diverse local contexts, identifying common characteristics can support the development of shared yet context-sensitive solutions. Through collaboration, cities do not need to reinvent the wheel; transportation strategies that prove effective in one city may be successfully adapted to others with similar profiles. Equally important, recognizing similarities can help cities avoid repeating costly mistakes made elsewhere, leading to more efficient use of public resources and faster implementation of effective measures.

While a wide range of studies have classified cities based on road geometry (Badhruudeen et al., 2022), transport network topology (Tundulyasaree, 2019; Yamaoka et al., 2021), land use composition (Puissant & Eick, 2024), and mobility patterns (Coenegrachts et al., 2024), most of these approaches focus on a single domain. As a result, they provide valuable but partial insights into the structure and performance of urban transport systems.

There remains a lack of integrated methods that account for the interdependencies between network structure, activity distribution, and mobility. The interaction between these domains, such as how activity locations influence travel demand or how population dispersion and mobility affect network performance, is often not considered. This study addresses that gap by combining indicators from road topology, population and facility density, and mobility. In doing so, it enables a more holistic classification of urban transport systems across European cities and supports a deeper understanding of how urban transport networks are shaped.

This research contributes to the field by:

- Integrating multiple domains (road network, activity distribution, mobility) into a single clustering framework, offering a more comprehensive and holistic perspective on urban transport systems.
- Demonstrating the value of applying multiple clustering methods (K-Means, K-Medoids, Ward’s Method) to assess the robustness and interpretability of resulting city classifications.
- Providing a systematic methodology that combines public data sources, dimensionality reduction, and clustering evaluation to support cross-city comparisons and the identification of transferable strategies.

The objective of this research is to identify relationships between a broad set of urban characteristics that influence the functioning of transportation systems. The analysis focuses on European cities, where consistent and accessible data sources are available. By examining how these characteristics relate to one another, the cities are grouped using three clustering methods. Both differences between clusters and similarities within clusters are analyzed to understand the underlying structure of potential groupings. The selected characteristics cover five domains: road topology, population distribution, economic activity, mobility, and congestion.

The societal relevance of this research lies in its potential to support knowledge transfer between cities. If cities within the same cluster face similar challenges and share comparable characteristics, they may benefit from adopting successful transportation strategies from one another. In addition, cities can also look beyond their own cluster to identify better-performing cities in other groups, to explore which characteristics account for these differences and identify possibilities for improvement. Therefore, this methodology provides a practical foundation for identifying context-sensitive solutions across diverse urban environments, and may also be relevant for cities beyond Europe.

To guide this research, the following research question and sub-questions are formulated.

Research Question

How can the application of multiple clustering methods reveal distinct groups of European cities based on road network, activity distribution and mobility characteristics?

Sub-Questions

- Which indicators most effectively distinguish cities in terms of road topology, activity distribution, and mobility characteristics?
  How can these indicators be quantified and compared across European cities?
- What relationships exist between the included indicators, and what do these relationships reveal about underlying urban dynamics?
- Which clustering methods are most suitable for grouping European cities based on these indicators, and what are their respective advantages and limitations?
- How consistent are the clustering outcomes across different methods, and how can their similarities and differences be interpreted?

In Chapter 2, the literature foundation of this research is presented, including descriptive characteristics of urban areas and an overview of applied clustering methods. Chapter 3 details the methodology used to define urban boundaries, quantify the selected indicators, and apply correlation and clustering techniques. Chapter 4 presents the characterization results of the included cities and serves as the foundation for the clustering analysis. The clustering results, their interpretation, and underlying correlations are discussed in Chapter 5. Finally, Chapter 6 offers a reflection on the findings and limitations of the research, while Chapter 7 presents the main conclusions and provides recommendations for future work.

2
Literature Research

This chapter reviews the relevant literature that forms the foundation for the research presented in this report. Section 2.1 begins by defining the boundaries of an urban area, with a focus on the European context. Section 2.2 focuses on the urban transportation network and its relevant characteristics for the analysis. Section 2.3 explores the clustering methods commonly used in urban studies and how they are applied.

2.1. Urban Boundaries

Defining urban boundaries is a common challenge in urban research, particularly in the context of transportation analysis. Traditional administrative boundaries, while available and widely used, often fail to capture the full extent of urban functions. This is especially true in cases where commuting patterns and the continuous built environment extend beyond municipal borders. To address this, different methods have been developed to better align urban area borders with actual patterns of human activity and infrastructure use.

The challenge of defining urban boundaries objectively is further illustrated by the work of Chen et al. (2022), who argue that appropriate boundary definitions can vary by context. Their clustering-based approach identifies a characteristic spatial scale by analyzing how the number of urban clusters changes with different search radii. This adaptive method emphasizes spatial connectivity in comparison with fixed administrative or demographic thresholds, highlighting the diverse and context-dependent structure of urban form.

A common approach to defining urban areas is based on population density. These methods typically identify urban areas by applying thresholds for inhabitants per square kilometer and additional rules for the built environment continuity. For example, the European Commission's Degree of Urbanisation (DEGURBA) classification system identifies three types of areas: high-density urban centers, urban clusters, and rural grid cells (Eurostat, 2023). A high-density cluster consists of adjacent 1 km^2 grid cells with at least 1,500 inhabitants/km^2 and a minimum population of 50,000. Urban clusters include cells with at least 300 inhabitants/km^2 and 5,000 total population, while rural grid cells fall outside these categories.

Population-based methods such as DEGURBA are complemented by conceptual definitions of "urbanness." Weeks (2010) argues that urban areas should not be understood solely in demographic terms, but rather as place-based constructs shaped by population size and density, economic organization, and the transformation of natural or agricultural land into the built environment. Similarly, Mela (2014) describes urban areas as continuous settlements with higher population densities than their surroundings, while acknowledging that national interpretations vary.

These conceptual perspectives complement data-driven classification systems and offer valuable insight into the diversity of urban definitions across different contexts. For international comparison and effective monitoring of Sustainable Development Goal 11[1], UN-Habitat (2020) proposes a standardized global framework that combines built-up density with population-based criteria. Their dual classification, Urban Extent and DEGURBA, is summarized in Table 2.1.

Table 2.1: Constraints for Urban Extent and DEGURBA classifications (UN-Habitat, 2020).

Urban Extent	DEGURBA
• Urban built-up area: pixels where the walking distance circle has a built-up density greater than 50%. • Suburban built-up area: pixels where the walking distance circle has a built-up density between 25%–50%. Includes subdivided land, whether built-up or not. • Rural built-up area: pixels where the walking distance circle has a built-up density less than 25% and that are not on subdivided land.	• High-density cluster: adjacent 1 km^2 grid cells with ≥1,500 inhabitants/km^2 and a minimum population of 50,000. • Urban cluster: adjacent 1 km^2 grid cells with ≥300 inhabitants/km^2 and a minimum population of 5,000. • Rural grid cell: all other grid cells.

In the Dutch context, a more detailed approach is adopted by Statistics Netherlands (CBS), which classifies urban areas based on address density, the number of addresses within a one-kilometer radius around each address point (Centraal Bureau voor de Statistiek, 2023). Areas are divided into five categories: very strongly urban (≥2,500 addresses/km²), strongly urban (1,500–2,500), moderately urban (1,000–1,500), weakly urban (500–1,000), and non-urban (<500). While developed to reflect residential patterns, this method can also capture zones of economic activity, such as office hotspots, although these are generally less densely concentrated. Address density therefore functions not only as an indicator of population distribution but more broadly as a measure of the spatial distribution of human activity.

Building on the notion that the physical layout of a city underlies its functional dynamics, other studies have proposed using intersection density to define urban areas. A good example is the method developed by Borruso (2003), who applied Kernel Density Estimation (KDE) to compare address density and intersection density within an Italian municipality. The results showed that intersection density closely aligns with patterns in the built environment while adding a structural dimension by focusing on the road network. Since intersections are present in both residential and mixed-use areas, this method may better capture urban form in cities where living and working functions are spatially interrelated. While the exact applicability depends on the research focus, an intersection-based approach offers strong relevance for transport-oriented studies.

Table 2.2: Comparison of population, address, and intersection-based urban boundary methods.

Criteria	Population-based	Address-based	Intersection-based
Captures	Population density	Human activity (residential + work)	Urban structure and connectivity
Strengths	Standardized, widely used	High resolution, context-specific	Relevance for mobility analysis
Limitations	Less sensitive to land use mix	May underrepresent work zones	Depends on network data quality

As summarized in Table 2.2, population-based, address-based and intersection-based methods each emphasize different aspects of urban structure. Population density offers consistency for comparison, address density reflects residential and economic activity at a finer scale, and intersection density captures the physical layout of the street network, making it especially suitable for transport and mobility studies. Overall, each method captures different aspects of urban structure, and the choice of boundary should therefore be guided by the specific research objective and context.

[1] Sustainable Development Goal 11: Make cities and human settlements inclusive, safe, resilient and sustainable (United Nations Department of Economic and Social Affairs, 2012).

2.2. Urban Transportation System & Network Structure

An urban transport network encompasses the interconnected system of roadways, railways, and public transport infrastructure that facilitates the movement of people and goods within and between urban areas (Loo, 2009). These networks include both physical infrastructure, such as streets, highways, and rail lines, and service networks, such as public transportation schedules and routes. The effectiveness of these networks is essential in determining the level of urban accessibility, economic productivity, and overall mobility (Lin & Ban, 2013). In addition, urban road network topology is closely linked to a city's socioeconomic structure, with spatial constraints shaping both network design and mobility patterns. Considering these interdependencies in urban planning can improve accessibility and contribute to more efficient urban systems (Tsiotas & Polyzos, 2017).

Urban transport networks operate at multiple spatial and functional levels, requiring careful planning and management. Their development is shaped by historical urbanization patterns, economic activities, and governance structures (Rodrigue & Ducruet, 2016). Transport networks must balance efficiency, cost, and accessibility, all of which are constrained by land availability, financial resources, and mobility demand. These networks also develop in response to population growth, technological advancements, and environmental challenges, leading to continuous changes in urban mobility infrastructure.

The study of urban transport networks builds on principles from network science, a field rooted in graph theory and complex systems analysis (Ding et al., 2019; Ortman et al., 2020). This perspective helps to conceptualize cities as interconnected systems, where relationships between nodes (intersections or transit stops) and edges (roads or rail lines) shape the structure and functioning of the network. Network science enables researchers to evaluate urban transport systems in terms of connectivity, efficiency, and resilience (Porta et al., 2006).

These analytical tools offer insight into how network structures affect congestion, accessibility, and flow distribution. Ding et al. (2019) describe transport networks as self-organizing systems shaped by spatial constraints, infrastructure investments, and evolving mobility demands. Their work highlights three key challenges: maintaining a balance between efficiency and redundancy, ensuring equitable accessibility, and improving resilience to disruptions. A well-designed network must therefore balance competing priorities, as overly direct networks may result in bottlenecks, while excessive redundancy can increase costs without necessarily improving performance.

In this context, the topology of an urban transport network plays a fundamental role in determining how efficiently people and goods move through the city. Topological characteristics influence not only route availability and travel distances but also the robustness of the network under stress, making them critical for effective transport planning.

For measuring efficiency, the meshedness coefficient assesses how grid-like a network is. Networks with high meshedness values tend to offer multiple route options, reducing travel times and improving efficiency in case of disruptions (Strano et al., 2013). It is also considered a measure of redundancy. The average shortest path length offers an alternative efficiency metric, representing the mean number of steps required to travel between all pairs of nodes in the network. Shorter average path lengths imply more direct and efficient connections, reducing overall travel time and improving accessibility (Feng et al., 2022). However, in sprawling urban networks, longer shortest paths can result in increased reliance on a few key corridors, leading to higher congestion levels due to these bottlenecks. To compare cities of different sizes, the network efficiency ratio is introduced, defined as the ratio between the average shortest path length and the network diameter (the longest shortest path in the network) (Tsiotas & Polyzos, 2015). A low efficiency ratio suggests sprawling or tree-like structures, often found in suburban or fragmented networks, which limit accessibility. A high ratio, by contrast, points to a well-connected, grid-like layout with multiple redundant paths and improved flow.

Strano et al. (2013) also describe the average node degree, which refers to the average number of edges connected to a single node, offering insights into how interconnected a network is. Higher node degree values indicate a denser and more connected network, typically seen in grid-based networks that provide multiple route options (Louf & Barthelemy, 2014). A low degree, however, suggests that traffic is funneled into higher-level roads more quickly, which can increase congestion at major intersections. Strano et al. (2013) further observe that such low-degree networks are often tree-like in structure, reflecting self-organized or organically grown urban forms.

Another important metric is betweenness centrality, which measures how many shortest paths between node pairs pass through a given node (Feng et al., 2022). It quantifies the structural importance of nodes by identifying those that frequently act as 'bridges' in the network. A node with high betweenness centrality plays an important role in overall connectivity, its removal would significantly disrupt travel between different parts of the city (Cardillo et al., 2006). This makes betweenness centrality a valuable measure of both network efficiency and robustness.

Research comparing urban road networks shows that cities with a highly centralized betweenness distribution are more prone to congestion, as a small number of critical intersections handle a disproportionately large share of the traffic load (Strano et al., 2013). In contrast, cities with a more decentralized distribution of betweenness spread traffic more evenly across the network, reducing bottlenecks and improving resilience to disruptions (Louf & Barthelemy, 2014). Betweenness centrality is therefore also useful for identifying key transport corridors where interventions, such as rerouting strategies or infrastructure upgrades, can most effectively improve traffic flow and robustness (Tsiotas & Polyzos, 2015).

Thus, network topology serves as a fundamental component of network efficiency and robustness. It can also reveal congestion risks and illustrate trade-offs between connectivity, accessibility, and resilience in an urban transport network.

Public transport and active modes, primarily cycling and walking, play a key role in mitigating congestion risks and promoting sustainability within urban areas. Well-integrated public transport systems, such as those in Amsterdam and Copenhagen, feature multimodal hubs where metro, tram, and bus services are connected with cycling infrastructure to enable smooth mode transfers. Research by Technical University of Munich (2024) on bicycle integration in public transport highlights the benefits of bike-sharing stations at these hubs, which improve the first- and last-mile segments of trips and encourage public transport usage.

Moreover, cities that invest in cycling infrastructure and pedestrian-friendly design tend to achieve higher shares of active mobility, helping reduce dependency on private vehicles. The structural properties of public transport networks, such as transfer points and network centrality, can be studied using graph theory (Pu et al., 2022), while observed modal shares provide valuable insight into how active and public transport modes contribute to overall network performance.

2.3. Characterization & Clustering Cities

Cities differ not only in their physical layout but also in how these characteristics influence mobility patterns and transportation network performance. Characterizing cities based on their structural, functional, and behavioral attributes is important to understand broader urban dynamics. By grouping cities with similar features, clustering methods help identify shared patterns and distinguish urban typologies. Existing studies have applied these methods to systematically analyze spatial features, such as facility distribution and urban form, and explore how these influence the network performance.

The spatial distribution of population and facilities plays a central role in shaping a city's identity and transportation system performance. This dispersion directly influences residents' travel behavior, including their mode choices for commuting, shopping, and leisure, and thus affecting network efficiency. The distribution of facilities is therefore essential for both economic activity and transport choices (Um et al., 2009). Commercial facilities tend to concentrate in densely populated areas, where high foot traffic boosts accessibility. While these hubs support regional productivity, it can also increase congestion levels in central urban zones (Temeljotov Salaj & Lindkvist, 2020). Public facilities, such as hospitals and schools, follow different spatial patterns. Their location is typically regulated by government equity guidelines to make sure that all inhabitants have access. However, due to higher demand, they too are often concentrated in central urban areas. According to Um et al. (2009), the density of public facilities increases more slowly than that of commercial ones.

In suburban, low-density environments, limited facility availability leads to longer travel distances, reinforcing car dependency and increasing transport costs. By contrast, compact cities with dense service networks support shorter trips and greater use of the public transport system and multimodal mobility (Rodrigue & Ducruet, 2016).

Urban form, particularly the distinction between monocentric and polycentric structures, further influences transport efficiency and network resilience. Monocentric cities are organized around a dominant economic hub, often the central business district (CBD), where jobs and commercial activity are highly concentrated. These cities often have radial transport networks leading to the center. While this improves access to this center, it also risks overloading main corridors and increases travel distances for suburban areas (Lemoy, 2024).
In contrast, polycentric cities feature multiple economic sub-centers, which distribute activity and mobility demand across a greater area. This structure helps reduce congestion at the urban core and improves regional accessibility (Veneri, 2014). Sun et al. (2013) show that polycentricity can emerge from planned decentralization or organic urban expansion, as observed in Shanghai, where additional employment centers eased traffic bottlenecks. Similarly, Fu et al. (2017) find that in Wuhan, sub-centers helped balance commuting flows and reduced congestion in the center. However, the success of polycentric structures depends on sufficient investment in transport infrastructure and well-integrated multimodal systems (Veneri, 2014).

Clustering methods are widely used to group cities with similar structural or functional features, offering insights into shared challenges and potential solutions. Such classifications support comparative analysis and inform urban planning strategies.

Tundulyasaree (2019) applies graph-theory indicators to cluster rail-bound public transport networks in cities across four continents, primarily in Europe. The analysis uses metrics such as betweenness and closeness centrality, alpha index, clustering coefficient, and network efficiency. K-means clustering is then applied to group cities based on the topology of their PT infrastructure, without considering service frequency or capacity. Additionally, hierarchical clustering is used to validate the results and explore the underlying structure of the clusters.

A road geometry-based approach is presented by Badhruudeen et al. (2022), who analyze the street networks of the world's 80 most populous cities. They identify a linear relationship between the number of nodes and links, and define five distinct road network typologies:

- **Gridiron Cities:** High proportion of 90-degree street angles, typically found in planned, orthogonal grid layouts.
- **Long Link Cities:** Dominated by long, straight road segments, often observed in Chinese cities designed to optimize long-distance accessibility.
- **Organic Cities:** Characterized by short links and irregular angles, typically emerging from unplanned or historical development.
- **Hybrid Cities:** Combining short and long links with a balanced distribution of 90-degree angles.
- **Mixed Cities:** Incorporating features from multiple typologies without a dominant geometric structure.

Their findings suggest that road network morphology is shaped by a combination of geographic context, historical development, and planning policy. Most European cities fall within the Organic category, while Chinese cities tend to exhibit Long Link characteristics.

Comparably, Yamaoka et al. (2021) use betweenness centrality to classify urban street networks in 30 European cities. This research is based on road data obtained from OpenStreetMap using a tool called OSMnx. By analyzing local betweenness, the study distinguishes street segments that facilitate long-distance travel and those that support local pedestrian movement. Critical connections tend to be concentrated along major streets, while pedestrian-oriented areas cluster in central business districts (CBDs) and historic centers. This type of classification helps identify congestion-prone corridors and supports strategies aimed at improving walkability.

A broader spatial perspective is offered by Puissant and Eick (2024), who applies prototype-based clustering methods to examine city composition based on land use and building typologies. Their study identifies homogeneous areas, such as residential, commercial, or industrial zones, and maps how these spatial clusters vary between cities. This method provides insight into the internal organization of urban areas and how land-use patterns influence broader urban form.

Mobility behavior is the focus of Coenegrachts et al. (2024) research, who classify 311 European cities using both K-means and latent class clustering. Their analysis, centered on shared mobility services, reveals that cities with a rich supply of shared mobility options tend to have stronger economic potential, while smaller cities often present fragmented or underdeveloped shared mobility markets. The study provides insights into how urban transport policies and infrastructure shape mobility behavior. It not only highlights inequalities in mobility options but also underscores the importance of aligning urban transport policies with socioeconomic patterns. By doing so, cities can create more inclusive, efficient, and adaptive mobility systems that support more livable and accessible urban environments.

To conclude, clustering methods provide valuable tools for grouping cities based on a wide range of urban characteristics, including transport network structure, facility distribution, and mobility behavior. Approaches based on graph theory, road geometry, shared mobility, and spatial composition each contribute a distinct perspective. The choice of method depends on the specific research objective and the availability of relevant data. Collectively, these methods support comparative urban research by revealing underlying patterns and structural similarities across different contexts.

3 Methodology

This chapter presents the methodological framework for characterizing and clustering European cities. It presents the reasoning behind the approach and outlines each step of the process. Section 3.1 discusses the city selection criteria, the list of included cities, determining their boundaries, and the data sources. Section 3.2 introduces and defines all included indicators. Section 3.3 explains how certain indicators are calculated to characterize the urban transport networks. Finally, Section 3.4 describes the indicator preparation and the clustering methods applied. Figure 3.1 shows a flow chart of the methodological process.

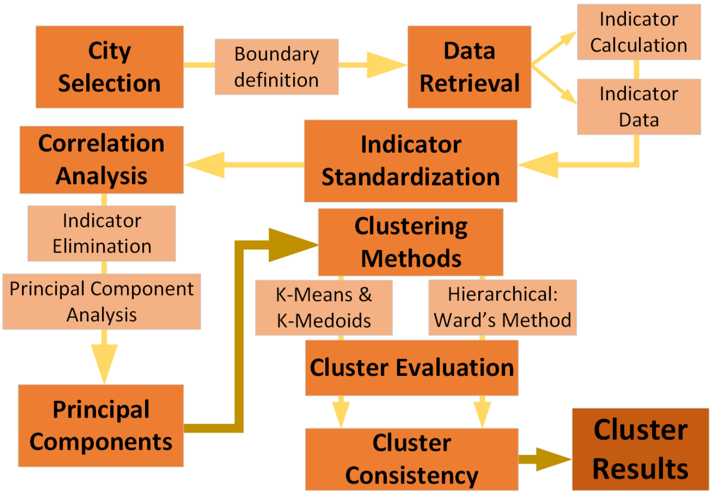


Figure 3.1: Methodological flowchart.

3.1. City selection & Urban boundaries

To evaluate the effectiveness of the clustering methods, it is necessary to obtain data from different urban transport networks. While one option was to simulate various networks with customized attributes and demand characteristics, the decision was made to focus on real cities in Europe to ensure the applicability of the results to practical, real-world contexts.

Based on the definition of a city outlined in Chapter 2, there are an estimated 10,000 cities globally (Scruggs, 2020). As shown in Figure 3.2, the European continent is densely populated. According to the definition used by the European Union, there are 828 cities in Europe[1] (Dijkstra & Poelman, 2012). These agglomerations vary in size, region and network structure. Since this research considers urban characteristics for different domains, data availability is essential.

[1] This number is based on population data for the EU, United Kingdom, Iceland, Norway, Croatia, and Switzerland.

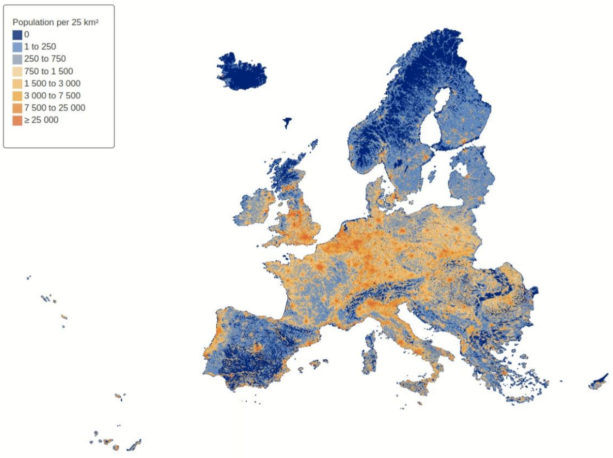


Figure 3.2: Population distribution over Europe (Eurostat, 2022).

Compared to other continents, Europe offers substantial data availability. However, due to the diversity of countries, there are multiple data collection methods and differing definitions of characteristics. This research relies on data sources where the data is collected or approved using consistent methodologies. While various sources will be discussed at the end of this section, it is important to note that mobility data, including modal share and car ownership, is the most difficult to obtain. Modal share data is often unavailable or not publicly accessible. Even when mobility data is available, it is often collected using inconsistent methodologies, which complicates cross-city comparisons.

Therefore, city selection is based on the report by EMTA (2024), which provides public transportation data for 35 European cities. For cities without modal share data for 2023 in this report, the earlier edition by EMTA (2022), containing data from 2020, is used instead. Since congestion levels reported by TomTom International BV (TomTom, 2024b) are also included in the clustering analysis, cities must have both congestion and mobility data available.
Athens and Porto are excluded due to the absence of recent modal share data in both EMTA reports, while Belgrade is excluded due to missing congestion data. The remaining 32 cities, located across 20 countries, are listed in Table 3.1.

Table 3.1: List of the 32 included cities with their respective countries.

City	Country	City	Country
Amsterdam	Netherlands	Barcelona	Spain
Berlin	Germany	Bilbao	Spain
Birmingham	United Kingdom	Brussels	Belgium
Bucharest	Romania	Budapest	Hungary
Copenhagen	Denmark	Frankfurt am Main	Germany
Helsinki	Finland	Krakow	Poland
Lisbon	Portugal	London	United Kingdom
Lyon	France	Madrid	Spain
Manchester	United Kingdom	Oslo	Norway
Palma de Mallorca	Spain	Paris	France
Prague	Czech Republic	Rotterdam	Netherlands
Sofia	Bulgaria	Stockholm	Sweden
Stuttgart	Germany	Thessaloniki	Greece
Toulouse	France	Turin	Italy
Valencia	Spain	Vienna	Austria
Vilnius	Lithuania	Warsaw	Poland

As explained in Section 2.1, city boundaries can be defined in different ways. Because this research focuses on urban transport networks, identifying representative urban boundaries is essential. While administrative boundaries are suitable in some contexts, Figure 3.3 shows that they can either exclude important network components, such as nearby cities that are functionally part of the urban transport system, or include irrelevant areas such as nature reserves or large bodies of water.

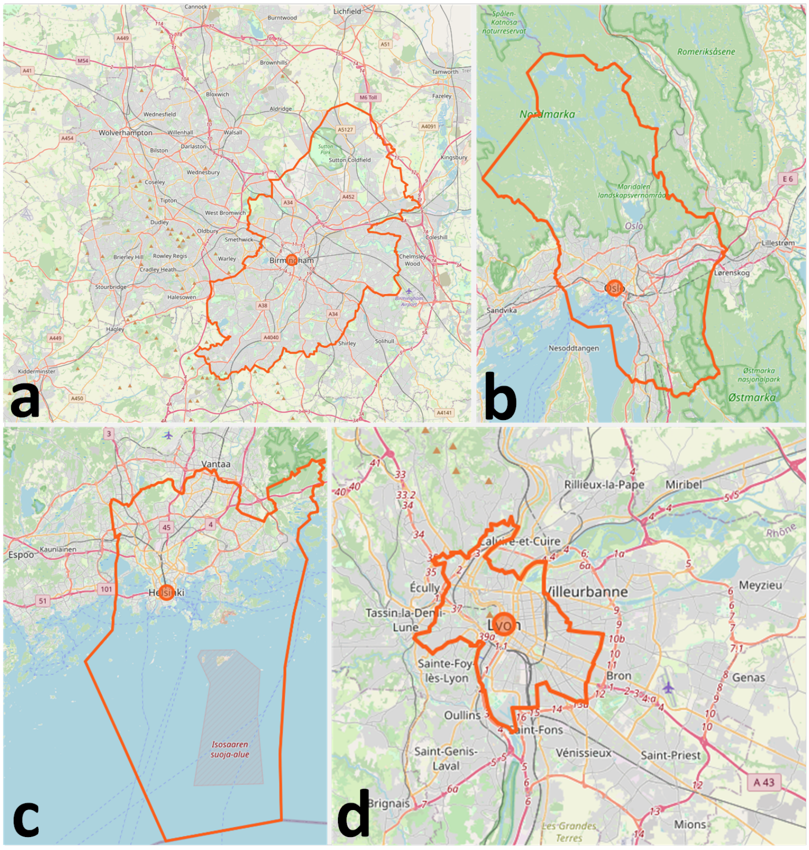


Figure 3.3: Administrative boundaries visualized for (a) Birmingham, (b) Oslo, (c) Helsinki, and (d) Lyon (OpenStreetMap contributors, 2025).

To address these limitations, customized polygons were created for each city based on its road network. The boundary construction process, for which the the Python code is shown in Section E.1, consisted of the following four steps:

1. An initial polygon was manually drawn around the ring road of each city using Mapbox (2025).
2. A bounding box of 80 by 80 kilometers was generated around the initial polygon to capture all potentially relevant nodes.
3. Nodes located within the initial polygon were labeled as ”included.” A proximity check was then performed: any ”non-included” node located within a threshold distance of an ”included” node was relabeled as ”included.” This process repeated iteratively until no further nodes met the threshold condition.
4. Using all included nodes, a final boundary polygon was generated with the AlphaShape Python package (Bellock, 2021).

In general, the threshold distance was set at 200 meters, meaning that adjacent nodes within this distance were treated as part of the same urban area. For five cities (Barcelona, Madrid, Valencia, Stuttgart, and Manchester) a smaller threshold of 150 meters was applied. Specifically, Barcelona, Madrid, and Valencia included excessive peripheral areas under the default threshold, likely due to denser intersections in Spanish urban planning or differences in intersection data reporting for Spanish cities by OpenStreetMap contributors (2025). Similarly, Stuttgart and Manchester also contained areas not relevant to the functional transport network when the default threshold was used.

The AlphaShape method constructs a boundary around a set of points, allowing for concave shapes that capture the point locations. In this research, the points represent the nodes in the road network. This method is controlled by a single parameter, α, which determines how closely the boundary follows the nodes. Although α is dimensionless, it acts as a distance control because the nodes are projected to a metric system. Smaller α values produce detailed boundaries that wrap tightly around the outer nodes, preserving gaps, indentations, and separate structures where the network is sparse. Larger α values create smoother shapes by bridging gaps between nodes, merging nearby clusters into a single area. An α value of 50.0 was selected to balance alignment with the dense urban network structure and the inclusion of suburban areas located in small gaps.

The resulting polygons were saved in both the World Geodetic System 1984 (WGS 84) coordinate system, compatible with OpenStreetMap, and the Mollweide projection, required for extracting population data.

With the urban boundaries established, the next step involves identifying widely available data sources to describe the characteristics of the 32 cities. As previously mentioned, ensuring dataset uniformity across cities in different countries is a challenge. Moreover, the analysis relies on publicly available data, which may vary in quality because they are maintained by many users. To ensure the reliability of results, the selection of consistent and representative datasets is essential.

Datasets for all relevant characteristics, as listed in Table 3.2, have been identified. Road network data is sourced from OpenStreetMap, which is maintained by a global community of contributors. It offers extensive coverage and also includes facility data such as shops, amenities and offices (OpenStreetMap contributors, 2025). However, the open nature of this platform can also affect data reliability. Residential locations are also relevant, as they provide insight into population distribution and possible travel patterns, together with the facility data. Population density data is obtained from Copernicus, which provides high-resolution (100 by 100 meters) population estimates in the Mollweide projection (Schiavina et al., 2023). Mobility characteristics are sourced from the previously mentioned EMTA reports (EMTA, 2022, 2024). These reports cover all 32 cities and, for most of them, also provide car ownership per 1,000 inhabitants in the Public Transport Authority (PTA) area. Finally, congestion levels for the metro area of a city are obtained from the TomTom Traffic Index, which offers a large dataset covering over 500 cities globally, containing all those included in this research (TomTom, 2024b). Data collection methods are consistent across all cities, unless otherwise noted in Section 3.2.

Table 3.2: Urban characteristics, and their data sources including reference year.

Characteristic	Data Source	Year
Road Topology	OpenStreetMap	2025
Population	Copernicus	2020
Economic Activity	OpenStreetMap	2025
Mobility	EMTA	2023 / 2020
Congestion	TomTom (Traffic Index)	2024

3.2. Indicator Overview

This section provides a conceptual overview of all indicators used to characterize the selected cities. Indicators are grouped into five domains: road topology, population, economic activity, mobility, and congestion. Detailed explanations of their relevance and interpretation are presented here, while the corresponding calculations for certain indicators are shown in Section 3.3.

Since the cities vary in size, directly comparing the variation across cities could be misleading. To address this, the standard deviation is normalized by the mean for each city, resulting in the coefficient of variation (CV). This allows for meaningful comparison of variation across the 32 cities.

3.2.1. Road Topology

The topology of the road network captures its structural characteristics, including connectivity, efficiency, and robustness. These properties are important to understand how urban transport networks function in different contexts. In this research, five indicators are used to describe road topology: the mean node degree (k_μ), the coefficient of variation of node degree (k_{cv}), the efficiency ratio (R), the coefficient of variation of betweenness centrality ($C_{B,\mathrm{cv}}$), and the 95th percentile of betweenness centrality ($C_{B,95}$). To ensure comparability between cities of different sizes and structures, all analyses are conducted on directed graphs accounting for one-way roads in the network.

The node degree captures how many connections a given intersection has. In a directed graph, this includes both incoming and outgoing links. The mean node degree (k_μ) reflects the average number of route options available at intersections, providing insight into how well-connected a network is. A higher value suggests more routing flexibility for road users.

The coefficient of variation of the node degree (k_{cv}) complements this by indicating how evenly this connectivity is distributed across the network. A low k_{cv} suggests uniform intersection types across the city, whereas a higher value implies a mix of sparse and highly connected intersections. These indicators are calculated directly from the raw road network, defined by the polygons constructed in Section 3.1.

Before analyzing network efficiency and robustness, a key preprocessing step is required: node consolidation. When road network data is extracted from OpenStreetMap, intersections, especially those on highways, are often represented by multiple adjacent nodes, as shown in Figure 3.4a. This artificially inflates the number of nodes and misrepresents the role of intersections.

To correct this, closely located nodes are merged using the node consolidation method from the OSMnx Python package (Boeing, 2024). After this process, each intersection is ideally represented by a single node, as shown in Figure 3.4b. The consolidation threshold is set to 25 meters, which is a trade-off between the need to merge highway intersections and avoiding the unintended merging of distinct intersections in dense neighborhoods. Details of this process are provided in Section 3.3.

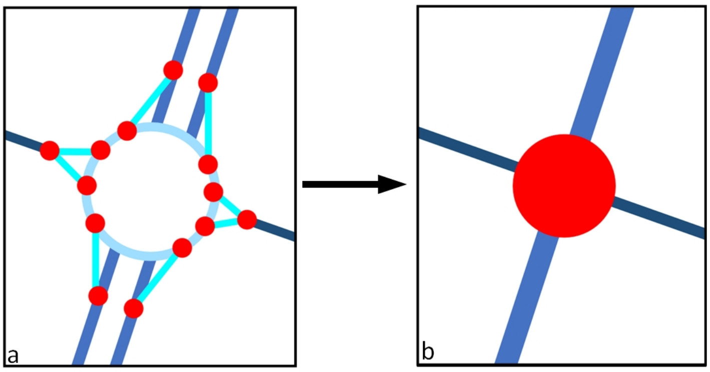


Figure 3.4: Node consolidation using the OSMnx Python package, showing (a) intersection nodes after network retrieval and (b) after node consolidation (Boeing, 2024).

For both the efficiency and betweenness indicators, weights must be assigned to the links in the road network. A straightforward approach might be to use the physical length of each segment, as it reflects the distance road users must travel. An alternative would be to use speed limits in combination with distance as weights, which would better approximate actual travel time and user preference.

However, both of these approaches present limitations in the context of OpenStreetMap data. Speed limit information is often missing from many road segments, making it unsuitable for consistent use. While segment length is consistently available, basing route choice solely on this metric can introduce bias. It does not reflect the preference of road users to favor higher-level roads with greater capacity, higher speed limits, and more comfort, regardless of their physical length.

In OpenStreetMap, higher-capacity roads such as motorways and trunks are typically represented by fewer, longer segments, whereas lower-capacity roads like residential streets consist of shorter segments. This effect is illustrated in Figure 3.5, where the longer but fewer segments on the primary road result in a lower link count than the many short segments on the secondary road. In a routing context where the number of links is used as the weight, the primary road is naturally favored, better reflecting real-world travel behavior.

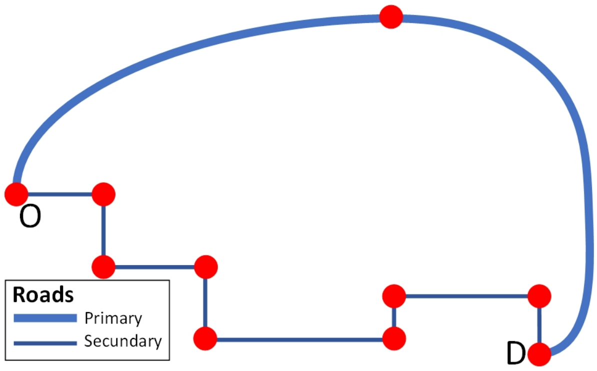


Figure 3.5: Use of the number of links as weight, illustrating that primary roads are favored over secondary roads.

Data on segment lengths by road type supports the logic behind this weighting choice. As shown in Table 3.3, roads with higher hierarchical tags (motorways, trunks) exhibit substantially longer average segment lengths compared to lower-level roads like residential streets and living streets. This structural pattern in OpenStreetMap reflects the functional road hierarchy and suggests that using the number of links as a routing weight can effectively approximate real-life routing preferences: longer, higher-capacity roads naturally consist of fewer segments and are more likely to be favored in shortest path calculations.

Because of limitations caused by missing or inconsistent speed limit data, this link-based weighting is used in both the efficiency ratio and betweenness centrality calculations. While not a perfect substitute for distance in combination with speed and capacity, it ensures consistency across cities and emphasizes the structural role of major roads in shaping network performance. The full analysis on segment characteristics is provided in Appendix C.

Table 3.3: Average segment length per road type in the consolidated network.

Road Type	Average Segment Length (m)
motorway	976.5
trunk	538.0
primary	240.8
secondary	205.8
tertiary	185.2
residential	157.9
living street	157.7

Network efficiency describes how easily areas within a city can be reached from one another. It is assessed by comparing the average shortest path (ASP) between all node pairs to the network's diameter (ND), defined as the longest shortest path in the network. The resulting efficiency ratio (R) is bounded between 0 and 1. Higher values indicate that most node pairs can be reached via relatively short routes compared to the longest shortest path between all node pairs, suggesting a well-connected network without the need for large detours. In contrast, lower values imply that many node pairs require disproportionately long paths, indicating inefficiency in the network.

Betweenness centrality quantifies how often a node lies on the shortest paths between other node pairs in the network. It reflects the importance of intersections in facilitating movement across the city. Nodes with high betweenness centrality are more likely to be involved in routing between origin–destination pairs, making them critical for the robustness and connectivity of the network. In this research, two indicators are included to capture aspects of the betweenness centrality.

The first is the coefficient of variation of betweenness centrality ($C_{B,\mathrm{cv}}$), which describes how evenly centrality is distributed across all nodes. A high $C_{B,\mathrm{cv}}$ suggests that a small number of nodes carry a disproportionately large share of network flow, indicating dependence on a few intersections. In contrast, a low $C_{B,\mathrm{cv}}$ reflects a more balanced distribution, where traffic load is spread more evenly throughout the network.
The second indicator is the 95th percentile of betweenness centrality ($C_{B,95}$), which highlights the centrality level of the most important nodes while minimizing the influence of extreme outliers. To enable comparison across cities, the $C_{B,95}$ values are normalized. This metric captures how dominant the top-ranking intersections are within a city's network.

Together, the two indicators provide insight into the robustness and, to a lesser extent, the connectivity of the road networks. As discussed in Section 2.2, high values of $C_{B,\mathrm{cv}}$ and $C_{B,95}$ indicate that a large share of shortest paths, and thus potential traffic flow, is concentrated on a small subset of nodes. This structural dependence increases the vulnerability of the network, as disruptions at these critical intersections can lead to detours and congestion. In contrast, lower values suggest a more evenly distributed network, where multiple alternative routes are available. This improves traffic dispersion and enhances the ability to maintain network functionality under stress.

These five road topology indicators (k_{μ}, k_{cv}, R, $C_{B,\mathrm{cv}}$, and $C_{B,95}$) were chosen because they capture complementary aspects of road network structure at the city-wide scale. The node degree indicators describe the connectivity and uniformity of intersections across the network, while the efficiency ratio reflects global accessibility by assessing whether large detours are necessary throughout the network. The two betweenness-based indicators measure network vulnerability and flow concentration, highlighting how traffic is distributed across intersections.

Other potential metrics mentioned in Section 2.3, such as meshedness and closeness centrality, were considered but ultimately excluded due to practical and interpretive limitations. Meshedness, which measures how grid-like a network is, is difficult to apply consistently across cities with different spatial extents and boundary definitions. It is also highly sensitive to local variations, as it can differ substantially between neighborhoods within the same city and is influenced by whether peripheral or fragmented areas are included. Closeness centrality, which indicates how close a node is to all others, becomes less meaningful in large networks because it is strongly affected by the city's size, shape, and node density. Even when normalized, it remains difficult to interpret across cities of varying scale. In contrast, the selected indicators are less sensitive to boundary effects, better reflect travel behavior, and offer greater interpretability for whole-network comparison. Together, they form a balanced set of metrics that capture key dimensions of network connectivity, efficiency, and resilience, supporting a more robust analysis across the 32 European cities.

3.2.2. Population Distribution

The mean population density (P_{μ}) represents the average number of inhabitants per square kilometer across the entire urban area. It is calculated using a 100 by 100 meter resolution grid from the Copernicus dataset and provides an overview of how densely populated the urban area is. Since the polygons used in this research are custom-defined based on road networks rather than administrative boundaries, this indicator captures residential land use both within and beyond official borders. It provides a baseline for interpreting spatial patterns in population distribution.
The coefficient of variation of population density (P_{cv}) reflects how unevenly the population is distributed across the urban area. A high P_{cv} indicates strong differences between densely and sparsely populated neighborhoods, suggesting a more centralized urban form. In contrast, a low value points to a relatively uniform population spread, with fewer pronounced differences between neighborhoods. This indicator is particularly useful for identifying spatial inequality in residential distribution and offers potential insights into travel patterns and transport demand within a city.
The 95th percentile of population density (P_{95}) captures the density value below which 95% of all grid cells fall, highlighting the upper range of population density while reducing sensitivity to extreme outliers. This indicator emphasizes the typical high-density neighborhoods within a city, such as central districts or urban cores. Compared to the mean, P_{95} offers a more focused view of the upper tail of the density distribution, providing insight into potential pressure points in the urban transport system.

Together, these three indicators provide a comprehensive image of population characteristics within each city. While the mean density captures the overall level of residential presence, the coefficient of variation and 95$^{\text{th}}$ percentile highlight differences in spatial distribution and local concentration. Understanding these patterns is important for interpreting possible transport demand, planning future infrastructure and comparing residential distribution across cities.

3.2.3. Economic Activity

To analyze economic activity within each urban area, this research uses location data from OpenStreetMap, which includes a wide range of facility types such as shops, amenities, offices and industrial sites. Among these, two categories are selected to represent activity that directly influences movement within the transport system: shops and offices. These locations correspond to where people work and where they purchase goods or services, both of which are common trip destinations.

Although OpenStreetMap includes many other facility categories, they often overlap with these two primary types. For instance, a location labeled as an amenity may also be categorized as a shop (e.g. pharmacies), while some industrial areas can be tagged as offices (e.g. factories). Because this overlap varies between cities and is subject to different documentation standards, selecting two clearly defined and consistent categories ensures a more reliable basis for comparison.

Several grid resolutions were tested to effectively capture the spatial dispersion of facilities across urban areas. Grids of 100 by 100 or 200 by 200 meters proved too fine-grained given the limited number of facility locations, while a 1 by 1 kilometer grid was too coarse to provide meaningful spatial differentiation, particularly for smaller cities.

The mean shop density (S_μ) reflects the average number of shop locations per square kilometer within the study area, based on a 500 by 500 meter resolution grid. This indicator provides a general sense of commercial activity and access to goods and services. While documentation practices in OpenStreetMap may vary across cities and countries, averaging across the full urban area allows for relatively consistent comparisons.

The coefficient of variation of shop density (S_{cv}) captures how unevenly shopping activity is distributed across the urban area. A high value suggests that shops are concentrated in specific zones, such as commercial centers or shopping streets, while other neighborhoods remain less active. A lower value indicates a more even spread of shopping locations throughout the city. This indicator highlights spatial differences in access to shops and supports the identification of possible movement patterns. Additionally, it is less sensitive to documentation inconsistencies, as it reflects the relative distribution of facilities rather than their total count.

The mean office density (O_μ) measures the average number of office locations per square kilometer across the study area. Like shop density, it is calculated on a 500 by 500 meter grid and reflects the overall presence of workplace-related activity within the urban area. Offices play a key role in shaping daily travel patterns, as they represent an important destination for commuting. By capturing the general intensity of employment locations, O_μ provides important context for interpreting commuting flows and employment intensity, and helps to compare how strongly office-based economic activity is represented in each urban area.

The coefficient of variation of office density (O_{cv}) describes how concentrated or dispersed office locations are across the city. A high value implies that office activity is clustered in a limited number of business districts, suggesting a central employment structure. In contrast, a low value indicates a more evenly spread of employment locations. This indicator is relevant for analyzing spatial accessibility to jobs, the decentralization of employment in offices and the potential for peak-hour travel flows to be distributed or concentrated. As with shop density variation, this indicator also helps reduce sensitivity to documentation inconsistencies, focusing instead on the relative distribution of offices.

Together, these four indicators provide a comprehensive perspective on economic activity within each urban area. By including both the intensity and spatial distribution of shops and offices, two key dimensions of travel demand are captured: consumer- and employment-related destinations. These indicators not only help to characterize the structure of economic activity but also offer valuable insights into their accessibility. The formulation ensures comparability across cities despite differences in documentation quality, making them a robust foundation for cross-city analysis.

3.2.4. Mobility Profile

Mobility behavior is a key characteristic of urban transport systems, reflecting how inhabitants move within their cities and what travel modes they rely on. To capture this, four indicators are included in this research: the modal share of motorized vehicles, public transport and active modes, along with the car ownership. These metrics provide insight into mode preference, and broader mobility culture across different urban areas.
The modal share data is obtained from the European Metropolitan Transport Authority (EMTA), which reports mobility statistics across Europe. The EMTA (2024) report distinguishes four categories: Motorized Vehicles (MV), Public Transport (PT), Cycling and Walking. The earlier EMTA (2022) report presents three categories: Motorized Modes, Public Transport and Active Modes (AM)[2]. To ensure comparability, all data is aligned into three standard categories: MV, PT and AM. When cycling and walking are reported separately, their sum is used for the Active Modes category. Motorized Modes and Motorized Vehicles are considered the same. This standardized approach allows consistent comparison across the 32 cities.

The modal share of motorized vehicles (M_{MV}) refers to the proportion of trips made using privately owned motorized transport, such as cars and motorcycles. This mode typically dominates in car-oriented cities and is often associated with higher levels of congestion and reduced space for other transport modes. A high value for M_{MV} tends to reflect urban structures with limited public transport availability, low walkability or high car accessibility.
The modal share of public transport (M_{PT}) captures the proportion of trips taken by bus, tram, metro or train. It reflects the accessibility, coverage and quality of a city's public transport system. High values for M_{PT} suggest well-developed networks offering frequent service and competitive travel times. This indicator is particularly useful for evaluating policy effectiveness in reducing car use and encouraging more sustainable mobility behavior.
The modal share of active modes (M_{AM}) combines walking and cycling trips, travel forms that are both sustainable and health-promoting. A high share of active modes often imply compact, mixed-use cities with safe, well-connected pedestrian and cycling infrastructure. This indicator not only reflects local travel preferences and mixed urban form but also points to possible health priorities. It highlights the extent to which cities support short-distance, non-motorized travel.

Car ownership (V_{own}) is measured as the number of cars per 1,000 inhabitants, based on the Public Transport Authority area where available. This indicator reflects mobility preferences and accessibility to alternative modes such as public transport. Higher ownership rates imply greater car dependency and more dispersed urban form. They can also reflect higher levels of household income or wealth. In contrast, lower rates suggest stronger support for sustainable transport, greater urban density or policies discouraging private vehicle ownership and use. This indicator complements the modal share data by offering additional insight into the structural reliance on private vehicles within each city, helping to distinguish between car ownership and actual car use.
For most cities, car ownership data is obtained directly from the EMTA (2022, 2024) reports. However, for Frankfurt am Main, Lisbon and Sofia, this information is not reported. In these cases, alternative sources are used to ensure a complete dataset: data for Frankfurt am Main is retrieved from Buehler et al. (2021), a national-level metric is used for Lisbon from da Costa (2024), and values for Sofia are based on estimates provided by Bergelings and Marchetti (2024).

The selected indicators offer a complete picture of mobility behavior across the 32 cities. By combining modal share data with car ownership levels, the analysis captures both daily travel preferences and longer-term mobility choices. This dual perspective highlights how transport systems are shaped by infrastructure and policies. To visualize the different mobility profiles of cities, a ternary plot is used to represent the relative shares of motorized vehicles, public transport and active modes. The standardized categories and data sources ensure consistency, enabling reliable cross-city comparisons of the overall mobility profile.

3.2.5. Congestion Level

The congestion level (CL) represents the average percentage increase in travel time due to traffic congestion. It is calculated by comparing actual travel times with those observed under free-flow

[2] The modal share data for Bilbao, Budapest, Thessaloniki and Vienna is obtained from the EMTA (2022) report.

conditions, as reported by the TomTom Traffic Index (TomTom, 2024a). This indicator shows how much longer trips take, on average, as a result of traffic delays, and is expressed as a percentage. The data is collected across all hours and days of the year, providing a consistent measure of overall traffic performance. Since it is collected at the metropolitan level, defined as the trip-dense region accounting for 80% of all recorded trips, it captures congestion across the full functional urban area rather than being limited to administrative boundaries or the city center.

This indicator is valuable for evaluating how well a transport network accommodates travel demand. A high CL suggests structural bottlenecks and reduced travel-time reliability. It may also point to broader systemic challenges, such as car dependence, limited modal alternatives or mismatches in infrastructure planning. In contrast, a low congestion level indicates more efficient traffic flow, improved accessibility and greater reliance on non-car modes. When analyzed alongside modal share and car ownership, CL adds a performance-based dimension to understanding the structure and functioning of the urban mobility systems.

The indicators presented in this section offer a structured basis for understanding differences in urban transport systems across the 32 cities. As summarized in Table 3.4, they span a range of infrastructural, spatial, behavioral and performance-related aspects. This includes the structural properties of road networks, population and economic activity distribution, mode choice preferences and congestion levels. By combining these dimensions, the indicators provide a balanced view of how urban areas are organized and how people move within them. This approach including multiple domains enables holistic comparisons between cities and forms the foundation for the clustering analysis.

Table 3.4: Overview of included indicators with their descriptions.

Indicator	Description
k_{μ}	Mean node degree, based on incoming and outgoing links per node (directed graph).
k_{cv}	Coefficient of variation (CV) of node degree, capturing variability in connectivity.
R	Efficiency ratio: average shortest path length divided by network diameter.
$C_{B,\mathrm{cv}}$	Coefficient of variation (CV) of betweenness centrality across all nodes.
$C_{B,95}$	95th percentile of betweenness centrality (normalized).
P_{μ}	Mean population density (inhabitants/km^2; 100×100m grid).
P_{cv}	Coefficient of variation (CV) of population density across the urban area.
P_{95}	95th percentile of population density.
S_{μ}	Mean shop density (locations/km^2; 500×500m grid).
S_{cv}	Coefficient of variation (CV) of shop density.
O_{μ}	Mean office density (locations/km^2; 500×500m grid).
O_{cv}	Coefficient of variation (CV) of office density.
M_{MV}	Modal share of motorized vehicles (cars and motorcycles; % of trips).
M_{PT}	Modal share of public transport (bus, tram, metro, train; % of trips).
M_{AM}	Modal share of active modes (walking and cycling; % of trips).
V_{own}	Car ownership, measured as vehicles per 1,000 inhabitants.
CL	Congestion level: average annual increase in travel time due to traffic delays (%).

3.3. Indicator Calculation

To perform the analysis for all 32 cities, several indicators require calculations. These calculations are based on the custom boundaries defined in Section 3.1. As a result, raw data must first be spatially matched to each polygon before computing individual indicator values. For road network indicators, this involves retrieving and processing OpenStreetMap data. Population indicators are derived from the Copernicus dataset, and economic activity indicators from OpenStreetMap location data. Mobility and congestion indicators are taken directly from existing sources and do not require further processing.

3.3.1. Road Topology

The topology of each road network is represented as a directed graph, where nodes represent intersections and edges correspond to road segments. The graph is retrieved using the OSMnx Python package (Boeing, 2024), which enables efficient extraction of road infrastructure based on geographic boundaries defined in the WGS 84 coordinate system. The road graph is retrieved using the following function:

```
osmnx.graph_from_polygon({City Polygon}, network_type="drive", simplify=True,
retain_all=False, truncate_by_edge=True)
```

Where:

- `network_type="drive"`: includes only public roads accessible to motorized vehicles.
- `simplify=True`: removes redundant nodes.
- `retain_all=False`: retains only the largest connected component within the polygon.
- `truncate_by_edge=True`: includes nodes outside the polygon if they directly connect to internal nodes.

To avoid distortions in shortest path and centrality calculations caused by closely spaced intersection nodes, common in OpenStreetMap highway intersection data, the network is first projected to a local metric coordinate system using `osmnx.project_graph({City Graph)}`, after which the built-in OSMnx function is applied to merge overlapping nodes through spatial buffering.
A buffer radius of 25 meters is applied around each node. If two buffer zones overlap, the corresponding nodes are merged. In theory, nodes within a maximum distance of 50 meters can be consolidated. While a larger buffer may improve consolidation of complex highway interchanges, it also risks merging distinct intersections in dense urban areas. A threshold of 25 meters is therefore selected to balance these considerations. After consolidation, the graph is projected back to WGS 84 to ensure consistent geographic referencing across the different analyses. The consolidation function is defined as:

```
osmnx.simplification.consolidate_intersections(G, tolerance=25, rebuild_graph=True,
dead_ends=True, reconnect_edges=True)
```

Where:

- `tolerance=25`: a per-node buffering radius of 25 meters.
- `rebuild_graph=True`: uses a topological, instead of geometrical, algorithm to identify close nodes.
- `dead_ends=True`: retains dead-end nodes in the network.
- `reconnect_edges=True`: reconnects the consolidated nodes and updates the edge lengths accordingly.

The consolidation process enhances the structural realism of the network, ensuring that intersections are properly represented and that shortest path and betweenness computations reflect actual network functionality.

Node consolidation is not applied for the node degree indicators, as it alters the spatial arrangement of nodes and their connecting edges. To preserve the original intersection structure and capture true connectivity, these indicators are computed based on the unconsolidated graph.

Node Degree Indicators

The node degree captures the number of connections each intersection has in the road network. Since the road graphs are treated as directed, both incoming and outgoing edges are counted separately. For a standard intersection with four two-way streets, this would result in a node degree of eight.

To assess the general connectivity of each network, the mean node degree (k_μ) is computed across all internal nodes. Nodes that are located outside the polygon are excluded from this calculation, even though they are included in the graph using `truncate_by_edge=True`. These external nodes are added to ensure that internal nodes retain all their connections, avoiding underestimation of internal node degrees at the polygon boundary. However, since the full set of connections for these external nodes is not available, including them in the analysis would introduce bias by consistently lowering the mean. To prevent this bias, only the internal nodes are used in the statistical calculations, while their connections to the external nodes are preserved. Equation 3.1 shows the formula.
In addition to the mean, the coefficient of variation of the node degree (k_{cv}) is calculated to quantify the relative variability in connectivity across the network. This is done by computing the standard deviation of node degree, using Equation 3.2, and then dividing it by the mean node degree as illustrated in

Equation 3.3. As with the mean, only internal nodes are considered in the calculation to avoid bias introduced by incomplete connectivity of external nodes.

$$k_\mu = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{3.1}$$

$$k_{\text{std}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (k_i - k_\mu)^2} \tag{3.2}$$

$$k_{\text{cv}} = \frac{k_{\text{std}}}{k_\mu} \tag{3.3}$$

Where:

- k_μ – Mean node degree of the internal nodes.
- k_{std} – Standard deviation of node degree.
- k_{cv} – Coefficient of variation of node degree.
- k_i – Degree of node i, measured as the number of incoming and outgoing edges.
- N – Total number of internal nodes considered.

A histogram of the node degree is also generated to visualize how connectivity is distributed across the network. This enables a more intuitive interpretation of the results by revealing whether most intersections share a similar number of connections, or if there is a wider spread including highly or sparsely connected nodes. These visualizations complement the statistical indicators by explaining aspects that are not directly captured by the statistics.

Efficiency Ratio

To evaluate the efficiency of the road network, the average shortest path length (ASP) is calculated. This represents the mean number of links traversed between all node pairs, based on the shortest paths in the network. However, since the cities included in this research vary considerably in size and structure, ASP values are not directly comparable.

To account for this, the ASP is normalized by the network diameter (ND), which is defined as the longest shortest path between any two nodes in the network. These metrics are computed using Equation 3.4 and Equation 3.5 respectively, with each link weighted equally, as introduced in Section 3.2. The values are calculated using the NetworKit library (Staudt et al., 2016), which enables efficient computation of both average shortest path and network diameter in large-scale directed graphs. The resulting ratio R is calculated using Equation 3.6.

$$\bar{d} = \frac{1}{N(N-1)} \sum_{i \neq j} d(i,j) \tag{3.4}$$

$$D = \max_{i,j} d(i,j) \tag{3.5}$$

$$R = \frac{\bar{d}}{D} \tag{3.6}$$

Where:

- N – Total number of nodes in the consolidated network.
- $d(i,j)$ – Shortest path between node i and node j, measured in number of links.
- $\bar{d}$ – Average shortest path length, referred to as ASP.
- D – Network diameter, referred to as ND.
- R – Efficiency ratio, defined as the ratio of ASP to ND.

Betweenness Centrality Indicators

To evaluate how traffic flow is concentrated within the network, betweenness centrality is calculated for each node. This metric captures how often an individual node lies on the shortest paths between all other node pairs, as defined in Equation 3.7. Here, $C_B(v)$ represents the raw betweenness centrality of node v. Shortest paths are based on the number of links between nodes, as discussed in Section 3.2. The centrality values are computed using the NetworKit library (Staudt et al., 2016), which applies Freeman normalization to ensure comparability across networks of different sizes.

$$C_B(v) = \sum_{s \neq v \neq t} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{3.7}$$

Two indicators are derived from the resulting distribution of node-level centrality values. The coefficient of variation ($C_{B,\mathrm{cv}}$) quantifies the relative variability of betweenness centrality, calculated as the standard deviation divided by the mean, as shown in Equation 3.8.
Additionally, the 95th percentile ($C_{B,95}$) characterizes the highest centrality values in the network while reducing the influence of extreme outliers. The Freeman normalization applied to this metric is shown in Equation 3.9.

$$C_{B,\mathrm{cv}} = \frac{C_{B,\mathrm{std}}}{C_{B,\mu}} \tag{3.8}$$

$$C_{B,95} = \frac{95^{\text{th}} \text{ percentile of } C_B(v)}{(N-1)(N-2)} \tag{3.9}$$

Where:

- $C_B(v)$ – Betweenness centrality of node v.
- σ_{st} – Total number of shortest paths between nodes s and t.
- $\sigma_{st}(v)$ – Number of those paths passing through node v.
- $C_{B,\mu}$ – Mean betweenness centrality of all nodes.
- $C_{B,\mathrm{std}}$ – Standard deviation of node betweenness.
- $C_{B,\mathrm{cv}}$ – Coefficient of variation of node betweenness.
- $C_{B,95}$ – Normalized 95th percentile of node betweenness.
- N – Total number of nodes in the consolidated network.

To support interpretation, a visualization of betweenness centrality per node is generated for each city. Node size and color both scale with individual centrality values, while the color scale is capped at the 95th percentile ($C_{B,95}$) to ensure visual clarity. These maps highlight critical intersections and provide spatial insight into the concentration of traffic flow.

3.3.2. Population Indicators

Population density values are retrieved from the Copernicus dataset, which provides population estimates at a 100 by 100 meter resolution in the Mollweide coordinate system. As discussed in Section 3.2, this high-resolution raster allows for precise, grid-based population analysis that is independent of administrative boundaries. To ensure compatibility, the retrieved Mollweide polygons are used to extract values for each urban area.
For each overlapping raster file, population values are extracted from the grid cells that fall within the city polygon. Cells with no data are excluded, and valid values are multiplied by 100 to convert them from inhabitants per hectare to inhabitants per km^2.

From the resulting set of grid-cell values, three population indicators are computed to characterize both the magnitude and distribution of population across each city. These include the mean population density (P_μ), the coefficient of variation (P_{cv}) and the 95th percentile of population density (P_{95}).
The mean population density, defined in Equation 3.10, is computed as the average density across all 100 by 100 meter cells within the polygon. To assess the relative variability of population density, the coefficient of variation is calculated by dividing the standard deviation (Equation 3.11) by the mean

as shown in Equation 3.12. Finally, the 95th percentile (Equation 3.13) identifies the upper tail of the distribution, capturing the threshold beyond which only the most densely populated areas are located.

$$P_{\mu} = \frac{1}{N} \sum_{i=1}^{N} P_i \tag{3.10}$$

$$P_{\text{std}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (P_i - P_{\mu})^2} \tag{3.11}$$

$$P_{\text{cv}} = \frac{P_{\text{std}}}{P_{\mu}} \tag{3.12}$$

$$P_{95} = 95^{\text{th}}\text{percentile of } \{P_1, P_2, \ldots, P_N\} \tag{3.13}$$

Where:

- P_{μ} – Mean population density (inhabitants/km^2).
- P_{std} – Standard deviation of population density.
- P_{cv} – Coefficient of variation of population density.
- P_{95} – 95th percentile of population density.
- P_i – Population density of grid cell i in inhabitants/km^2.
- N – Total number of 100 × 100 meter grid cells within the polygon.

To complement the numerical indicators, a population density map is created for each city. For visualization purposes, the population data is projected to WGS 84 to ensure consistent geographic referencing with other visualizations. The color scale is capped at the 95th percentile to minimize the visual impact of outliers and improve interpretability. These visualizations support the interpretation of the statistical results by making density patterns visible, such as urban centers and polycentricity.

3.3.3. Economic Activity Indicators

The economic activity indicators are calculated using location data from OpenStreetMap, focusing on the density of shops and offices for each city. These locations serve as measures for commercial and work activity within urban areas. As outlined in Section 3.2, both land-use indicators are extracted within a 500 by 500 meter grid, enabling spatial comparison across cities.

The urban boundaries are projected to an appropriate local coordinate reference system using the UTM zone derived from the polygon centroid. This ensures accurate square grid creation, which is fitted over each polygon. Within each grid cell, the number of shops and offices is counted separately using the OpenStreetMap tags `"shop": True` and `"office": True`. The total number of facilities is then converted to densities (locations/km^2) using the area of each cell.

From the resulting gridded data, two indicators are calculated for both shops and offices: the mean density (S_{μ}, O_{μ}) which are shown in Equation 3.14, and the coefficient of variation (S_{cv}, O_{cv}) is calculated using the standard deviation (Equation 3.15) in Equation 3.16.

$$S_{\mu} = \frac{1}{N} \sum_{i=1}^{N} S_i \quad \text{and} \quad O_{\mu} = \frac{1}{N} \sum_{i=1}^{N} O_i \tag{3.14}$$

$$S_{\text{std}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (S_i - S_{\mu})^2} \quad \text{and} \quad O_{\text{std}} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (O_i - O_{\mu})^2} \tag{3.15}$$

$$S_{\text{cv}} = \frac{S_{\text{std}}}{S_{\mu}} \quad \text{and} \quad O_{\text{cv}} = \frac{O_{\text{std}}}{O_{\mu}} \tag{3.16}$$

Where:

- S_i – Shop density in grid cell i (locations/km^2).
- O_i – Office density in grid cell i (locations/km^2).
- S_μ, O_μ – Mean shop and office densities.
- S_{std}, O_{std} – Standard deviation of shop and office densities.
- S_{cv}, O_{cv} – Coefficient of variation (CV) of shop and office densities.
- N – Total number of 500 × 500 meter grid cells within the city boundary.

To support interpretation, a spatial density map is produced for each city, visualizing the density of shops and offices. The color scale is proportional to the maximum number of facilities per km^2, revealing commercial clusters and the intensity of economic activities in the urban areas.

3.4. City Characterization

Before performing the clustering analysis, the dataset for the 32 cities is first prepared to ensure that the selected indicators are both comparable and meaningful for grouping cities. This involves three steps: standardizing the 17 indicators to eliminate differences in scale in subsection 3.4.1, assessing correlations to uncover redundancy and relationships among variables in subsection 3.4.2, and applying Principal Component Analysis (PCA) to reduce dimensionality while retaining as much variation from the original data as possible in subsection 3.4.3.

3.4.1. Data Standardization

The indicators used in this research differ in both units and magnitude, which makes direct comparison difficult and potentially misleading. To ensure comparability and equal influence in the analysis, all indicators are standardized using Z-score normalization, as shown in Equation 3.17. This method transforms each indicator to have a mean of 0 and a standard deviation of 1, removing unit-based differences and allowing the analysis to focus purely on the variation in values across cities. As a result, each indicator contributes proportionally to the clustering and dimensionality reduction processes, regardless of its original scale.

$$Z = \frac{X - \mu}{\sigma} \tag{3.17}$$

Where:

- Z – Standardized indicator value,
- X – Original indicator value,
- μ – Mean of the indicator,
- σ – Standard deviation of the indicator.

Alternative normalization methods, such as Min-Max scaling and Robust scaling, were also tested. However, these approaches did not show substantially different results in terms of interpretability or clustering outcomes. Z-score normalization was selected as the preferred method because it preserves the relative structure of the data while avoiding the constraint of normalizing values within a fixed range. Although it is not specifically robust to outliers, it provides a consistent and interpretable transformation that performs well across indicators with varying distributions.

3.4.2. Correlation Analysis

After standardization, the relationships between indicators are examined to identify potentially redundant indicators and detect strong relationships. The Pearson Correlation Coefficient (r) is used as the correlation measure, as it quantifies the strength and direction of linear relationships between pairs of variables. This is particularly relevant because Principal Component Analysis (PCA), introduced in subsection 3.4.3, relies on such linear correlations to reduce dimensionality.

The Pearson coefficient is calculated using Equation 3.18. A value of $r = 1$ indicates a perfect positive linear relationship, $r = -1$ a perfect negative relationship, and $r = 0$ implies no linear association. It is important to note that this method is sensitive to outliers and does not capture non-linear relationships, which may affect interpretation when such patterns are present in the data.

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\sqrt{\sum(Y_i - \bar{Y})^2}} \tag{3.18}$$

Where:

- X_i – Individual data point of variable X,
- Y_i – Individual data point of variable Y,
- $\bar{X}$, $\bar{Y}$ – Means of variables X and Y.

The Spearman Rank Correlation Coefficient was also tested as a robustness check due to its resistance to outliers and ability to detect monotonic relationships, it was ultimately not used because it is not directly compatible with PCA and did not show significantly different correlations results.

To evaluate whether the observed correlations are statistically meaningful, a two-tailed significance test is performed using the critical t-value (t_{crit}). This test checks whether the strength of the correlation could possibly occur by chance given the null hypothesis, which assumes there is no correlation between two indicators. In this research, a significance level of $\alpha = 0.05$ is chosen. This means that there is a 5% chance of incorrectly rejecting the null hypothesis, so a correlation is considered statistically significant only if the probability of it occurring by chance is less than 5%.

The test statistic depends on the number of observations (n, the number of cities), which determines the degrees of freedom ($df = n - 2$). Two degrees of freedom are subtracted because the test first estimates the means of the two variables involved before evaluating their relationship. Based on these values, a critical threshold is derived for the absolute correlation coefficient. The calculations are shown in Equation 3.19 and Equation 3.20.

$$t_{\text{crit}} = t_{\text{ppf}}(1 - \alpha/2, df) \tag{3.19}$$

$$r_{\text{threshold}} = \sqrt{\frac{t_{\text{crit}}^2}{t_{\text{crit}}^2 + df}} \tag{3.20}$$

Where:

- α – Significance level,
- $df = n - 2$ – Degrees of freedom,
- t_{ppf} – Inverse cumulative t-distribution function,
- $r_{\text{threshold}}$ – Minimum absolute correlation required for significance.

For the dataset with 32 cities ($n = 32$), the resulting threshold is $r_{\text{threshold}} = 0.349$. All absolute correlation values above this threshold are considered statistically significant.

Before performing the PCA, a subset of indicators can be excluded to minimize redundancy and improve interpretability. This selection is based on the correlation analysis, where indicators with strong linear and explainable relationships are assumed to contain overlapping information. Although PCA accounts for such correlations by constructing orthogonal components, retaining multiple highly correlated variables may still result in disproportionate emphasis on specific information. This can reduce the diversity of urban characteristics captured in this research. By filtering out redundant indicators beforehand, the resulting components become more balanced and interpretable, improving their usefulness for the clustering process.

To determine which indicator to retain within each group of strongly correlated variables, several criteria are considered. Indicators are deemed strongly correlated when the absolute Pearson correlation coefficient exceeds $|r| \geq 0.65$. Among these, preference is given to indicators with greater variability across cities, as measured by their indicator-specific coefficients of variation, since they are more effective at distinguishing between cities. Indicators with known data limitations or possible methodological inconsistencies are given lower priority in favor of more robust alternatives that capture similar concepts. In the case of the modal share indicators, summing to 100%, only one is retained, as it mostly represents the inverse variation of the other modal shares.

3.4.3. Principal Component Analysis

To reduce the number of dimensions in the dataset while retaining the majority of its variation, a Principal Component Analysis (PCA) is applied. This preprocessing step streamlines the input for PCA, ensuring that the remaining indicators contribute unique and meaningful information to the dimensionality reduction and subsequent clustering. PCA transforms the original, potentially correlated indicators into a smaller set of uncorrelated components, known as principal components (PCs). Each PC is a weighted linear combination of the standardized indicators and captures a unique direction of variance in the dataset (Greenacre et al., 2022). The PCs are ordered such that the first component explains the greatest amount of variance, followed by the second, and so on. Equation 3.21 shows how the principal components are calculated based on the indicators.

$$\text{PC}_i = w_{i1}Z_1 + w_{i2}Z_2 + \cdots + w_{ip}Z_p \tag{3.21}$$

Where:

- PC_i – Score of the i^{th} principal component for a city,
- $Z_1, Z_2, \ldots, Z_p$ – Standardized values of the p included indicators,
- $w_{i1}, w_{i2}, \ldots, w_{ip}$ – Component weights (loadings) of the i^{th} PC.

The PCA is performed using the Singular Value Decomposition (SVD) algorithm, which decomposes the standardized data matrix and facilitates the calculation of principal components and their corresponding explained variance. The key steps in this process are summarized in Table 3.5.

Table 3.5: Steps for performing PCA using Singular Value Decomposition (SVD) (Greenacre et al., 2022).

Step	Description
Compute SVD	Decompose the standardized data matrix using SVD: $X = U\Sigma V^T$.
Compute Principal Components	Obtain principal component scores by projecting the standardized data onto the new axes: $X_{\text{PCA}} = U\Sigma$.
Compute Explained Variance	Determine the proportion of variance captured by each principal component using eigenvalues: $\lambda_i = \sigma_i^2$.

A scree plot is used to visualize the proportion of variance explained by each principal component and to help the selection of how many components to keep for meaningful clustering. The selection is based on the cumulative explained variance, with the included components required to capture at least 75% of the total variance in the indicators.

In addition, the contribution of each indicator to the included principal components is examined to ensure that its variance is sufficiently represented in the reduced space. This step also helps the interpretation of the principal components by identifying which indicators contribute the most to each PC. If an indicator contributes minimally across all retained components, this may indicate that the number of components should be reassessed.

3.5. Clustering

Clustering is used to group cities with similar characteristics based on the principal components derived from the earlier analysis. This study applies three methods: K-Means, K-Medoids, and agglomerative hierarchical clustering using Ward's method. These approaches are selected for their interpretability and suitability for relatively small datasets. The analysis begins with K-Means to identify initial patterns, followed by K-Medoids to assess robustness against outliers. Both methods are discussed in subsection 3.5.1. Ward's method is then used to examine the hierarchical relationships between clusters and individual cities, as introduced in subsection 3.5.2. Finally, the clustering outcomes are compared using the Adjusted Rand Index (ARI) and the Jaccard Similarity, which is elaborated on in subsection 3.5.3.

3.5.1. K-Means & K-Medoids Clustering

Partitioning-based clustering methods divide a dataset into a predefined number of clusters (k), assigning each data point to a group based on its distance to a central reference point, as defined by the specific method: K-Means or K-Medoids. K-Means is applied first due to its computational efficiency and widespread use. K-Medoids is then used to compare the clustering results, as it is more robust to outliers.

To determine an appropriate number of clusters (k), the Elbow Method is applied, evaluating the Within-Cluster Sum of Squares (WCSS) across increasing values of k. WCSS quantifies the compactness of clusters by summing the squared distances between each point and its assigned cluster center, as defined in Equation 3.22.

$$WCSS = \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2 \tag{3.22}$$

Where:

- k – Number of clusters,
- C_i – Set of data points in cluster i,
- x_j – Data point assigned to cluster C_i,
- μ_i – Centroid (or medoid) of cluster C_i,
- $\|x_j - \mu_i\|^2$ – Squared Euclidean distance between a data point and its cluster center.

As k increases, WCSS naturally decreases because data points are grouped into smaller clusters, reducing their average distance to the center. However, beyond a certain point, the additional improvement in compactness decreases. The infliction in the WCSS curve marks this transition, indicating a balance between simplicity and the ability to capture distinct patterns in the data. This point is used to determine an appropriate number of clusters for both K-Means and K-Medoids.

K-Means

K-Means clustering divides the dataset into k clusters by minimizing the variance within each cluster. This method iteratively assigns each data point to the nearest cluster centroid and updates the centroid positions based on the mean of the assigned points. The optimization problem is formulated in Equation 3.23.

In this study, K-Means clustering is applied using the k-means++ initialization strategy, which improves the selection of initial cluster centers by favoring points that are farther apart from each other. Compared to purely random initialization, this approach enhances convergence speed and reduces the likelihood of local minima being the end result. For each value of k (ranging from 2 to 10), the algorithm is executed 1,000 times using different random seeds.
Within each run, the `n_init` parameter is set to 50, meaning that 50 internal centroid initializations are performed and the one with the lowest WCSS is retained. From the 1,000 runs per k, the final clustering result is selected from the seed that produces the lowest WCSS overall, ensuring a stable and high-quality clustering outcome for further interpretation.

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{x_j \in C_i} \|x_j - \mu_i\|^2 \tag{3.23}$$

Where:

- k – Number of clusters,
- C_i – Set of data points in cluster i,
- x_j – Data point assigned to cluster C_i,
- μ_i – Centroid of cluster C_i,
- $\|x_j - \mu_i\|^2$ – Squared Euclidean distance between a data point and its cluster centroid.

K-Medoids

K-Medoids clustering is a more robust alternative to K-Means. Instead of computing centroids, it selects actual data points, called medoids, as the centers of clusters. This approach makes the method less sensitive to outliers and more suitable for datasets with irregular cluster shapes or when different distance metrics are required[3].

Unlike K-Means, which minimizes the variance around centroids, K-Medoids minimizes the total pairwise dissimilarity between data points and their assigned medoid. This tends to produce slightly less compact but more robust clusters, particularly when the dataset contains outliers. Cluster separation can be less sharp compared to K-Means, but the stability of the assignments increases. The optimization objective is shown in Equation 3.24.

$$\arg\min_{C} \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, m_i) \tag{3.24}$$

Where:

- k – Number of clusters,
- C_i – Set of data points in cluster i,
- x_j – Data point assigned to cluster C_i,
- m_i – Medoid of cluster C_i,
- $d(x_j, m_i)$ – Distance between a data point and its cluster medoid.

In this study, an exhaustive search is conducted to identify the optimal medoid configuration for each number of cluster counts (ranging from 2 to 8), ensuring the global optimum is found. Due to the algorithm's high computational complexity, especially as the number of possible combinations increases with k, a maximum of eight clusters is selected to ensure computational feasibility, even with parallel computing. For larger datasets or higher k values, heuristic methods such as Partitioning Around Medoids (PAM) are often used instead to reduce computational burden, but is not considered in this report. The elbow plot is used to evaluate the resulting optimal WCSS values across different k values.

3.5.2. Hierarchical Clustering

Hierarchical clustering constructs a nested hierarchy of clusters without the requirement of predefining the number of groups. In this research, agglomerative hierarchical clustering is applied. This method begins with each data point as an individual cluster, iteratively merging the two clusters that show the most similarity based on a chosen linkage criterion. Ward's method is specifically applied due to its effectiveness in minimizing within-cluster variance throughout the merging steps (Murtagh & Legendre, 2011).

Ward's linkage criterion selects clusters for merging by evaluating the increase in the total WCSS. The associated objective function, defined in Equation 3.25, calculates the variance increase (ΔE) that occurs when two clusters are combined.

$$\Delta E = \sum_{x \in C} \|x - \mu_C\|^2 - \left(\sum_{x \in C_1} \|x - \mu_{C_1}\|^2 + \sum_{x \in C_2} \|x - \mu_{C_2}\|^2 \right) \tag{3.25}$$

Where:

- ΔE – Increase in total within-cluster variance,
- C_1, C_2 – Clusters being merged,
- C – New cluster resulting from the merge,
- $\mu_C, \mu_{C_1}, \mu_{C_2}$ – Centroids of the clusters,
- $\|x - \mu\|^2$ – Squared Euclidean distance between a data point x and its respective cluster centroid.

To identify the optimal number of clusters, the dendrogram, a graphical representation illustrating how clusters are hierarchically merged, is ”cut” at various levels corresponding to different cluster counts

[3] For K-Medoids, Euclidean distance is used for consistency and comparability with the K-Means results.

(ranging from 2 to 10). The dendrogram clearly visualizes in which order and at what linkage distances which clusters merge, facilitating interpretation of the hierarchical structure. An example dendrogram is provided in Figure 3.6.

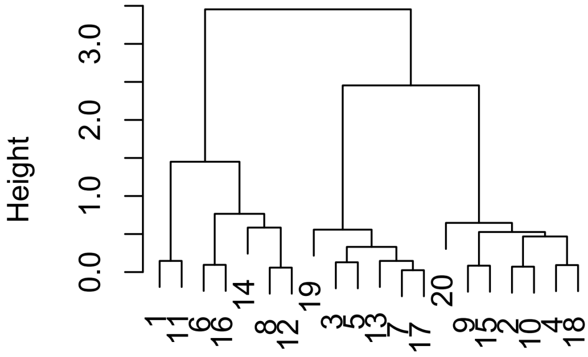


Figure 3.6: Example dendrogram demonstrating hierarchical city clustering (Murtagh & Legendre, 2011).

Overall, Ward's hierarchical clustering method provides a robust framework for exploring and interpreting hierarchical relationships within the data, particularly valuable when the optimal number of clusters is unknown or multiple number of clusters can be explored.

3.5.3. Cluster Evaluation

This subsection outlines the approach for evaluating the clustering results obtained from the three methods. First, the optimal number of clusters, potentially more than one, is determined using the silhouette scores for each number of clusters for each method. Next, the global agreement between clustering outcomes is assessed using the Adjusted Rand Index, followed by checking the stability of individual cities using the Jaccard Similarity.

Silhouette Score

To determine the optimal number of clusters for each clustering method, the silhouette score is used as an evaluation metric. This measure is applied to each method to assess the clustering quality across different values of k. For K-Means and K-Medoids, it complements the Elbow Method by providing a measure that accounts not only for compactness but also for the separation between clusters. In the case of Ward's Method, it offers a quantitative basis in addition to the interpretation of the dendrogram. The silhouette score evaluates how well each data point fits within its assigned cluster compared to its similarity to the nearest alternative cluster. It captures two aspects: the internal similarity of points within the same cluster and their separation from other clusters. For each observation i, the silhouette value $s(i)$ is computed as shown in Equation 3.26. The global silhouette score S_k for a clustering configuration with k clusters is calculated as the average of all individual scores, as shown in Equation 3.27.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}} \tag{3.26}$$

$$S_k = \frac{1}{n} \sum_{i=1}^{n} s(i) \tag{3.27}$$

Where:

- $a(i)$ – Average distance from point i to all other points in the same cluster (intra-cluster distance),
- $b(i)$ – Lowest average distance from point i to all points in the nearest neighboring cluster (inter-cluster distance),
- $s(i)$ – Silhouette value for point i, ranging from -1 to 1,
- n – Total number of data points,
- k – Number of clusters in the current configuration,
- S_k – Average silhouette score for the configuration with k clusters.

A silhouette value close to 1 indicates that the data point is well clustered, while a value near 0 suggests ambiguity between two clusters. Negative values imply that the point may have been assigned to the wrong cluster. The global silhouette score S_k is used in this study to identify the most optimal number of clusters per method.

To assess not only the optimal number of clusters within each method but also the relative quality of the methods themselves, the silhouette scores corresponding to the selected k values are compared across K-Means, K-Medoids, and Ward's Method. Since all methods are applied to the same standardized PCA-transformed data using Euclidean distance, these global silhouette scores provide a consistent basis for evaluating both clustering results and the quality of each method.

Adjusted Rand Index

The Adjusted Rand Index (ARI) offers a complementary perspective to the silhouette-based evaluation of clustering quality by measuring the overall agreement between clustering results from different methods. While the silhouette score assesses the internal cohesion and separation of clusters, the ARI evaluates all possible pairs of cities to determine whether each pair is assigned to the same cluster in both solutions. ARI values range from -1 (no agreement) to 1 (perfect agreement), with a value of 0 indicating a level of similarity no better than random chance.

Given two clustering results U and V of a set of n elements, let n_{ij} be the number of elements shared between cluster i in U and cluster j in V. Let $a_i = \sum_j n_{ij}$ be the total number of elements in cluster i of U, and $b_j = \sum_i n_{ij}$ be the total number of elements in cluster j of V. The ARI is then computed using Equation 3.28.

$$\text{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2} \Big/ \binom{n}{2}\right]}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2} \Big/ \binom{n}{2}\right]} \tag{3.28}$$

Where:

- n – Total number of data points (cities),
- n_{ij} – Number of cities assigned to both cluster i in U and cluster j in V,
- $\binom{n}{2}$ – Number of possible city pairs,
- ARI – Adjusted Rand Index score between two clustering methods.

The ARI is used here to compute pairwise similarity scores between the clustering outputs of K-Means, K-Medoids and Ward's Method. The resulting ARI values are presented in matrix form to provide a visual overview of the global agreement between methods.

Jaccard Similarity

While the ARI captures global agreement between clustering results, it does not reveal inconsistencies at the level of individual cities. To address this, a local robustness check is performed using the Jaccard Similarity between the neighborhoods of each city. A city's neighborhood is defined as the set of other cities that are assigned to the same cluster. The Jaccard Similarity compares these sets across methods, quantifying the consistency of a city's cluster membership across methods. For two neighborhood sets, the Jaccard similarity is calculated with Equation 3.29.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{3.29}$$

Where:

- A, B – Neighborhood sets (cities in the same cluster) for a specific city under two different clustering methods,
- $|A \cap B|$ – Number of cities shared in both neighborhoods,
- $|A \cup B|$ – Total number of unique cities across both neighborhoods,
- $J(A,B)$ – Jaccard similarity ranging from 0 (no overlap) to 1 (identical sets).

For each city, all pairwise Jaccard similarities between its neighborhood sets across methods are calculated and averaged. Cities with low average scores are labeled as inconsistently assigned, indicating that their cluster assignments vary notably using different methods. This neighborhood-based evaluation complements the ARI by identifying cities whose assignment within the cluster structure are unstable.

The combination of evaluation metrics introduced in this section is used to determine the most appropriate number of clusters. In particular, the silhouette score and Adjusted Rand Index (ARI) guide the overall selection by evaluating clustering quality and consistency across the three methods. These metrics are complemented by visual interpretation of the cluster configurations to assess whether the results correspond to meaningful geographical patterns. The Jaccard Similarity analysis serves as a more detailed metric, focusing on individual cities whose cluster assignments vary between methods. While it does not determine the number of clusters, it assists in interpreting inconsistencies within the cluster results.

4

Network Characterization

This chapter presents the results for the indicators introduced in Chapter 3, focusing on four selected cities: Amsterdam, Bilbao, Budapest, and Stockholm. These cities were chosen to reflect the variability in the dataset, covering different city sizes and geographical regions in Europe. Their results are discussed in Section 4.1, while the full dataset for all cities is provided in Appendix A. In Section 4.2, correlations between the seventeen indicators are analyzed to identify redundancies and support the exclusion of certain indicators from the clustering process. Finally, a Principal Component Analysis (PCA) is performed in Section 4.3 to remove remaining redundancy and determine the principal components on which the cities will be clustered.

4.1. Data Overview

To illustrate how the data is structured, Table 4.1, Table 4.2, and Table 4.3 provide an overview of all seventeen indicators introduced in Chapter 3. The four selected cities serve as illustrative examples, highlighting the diversity in urban characteristics observed across the full set of 32 cities.

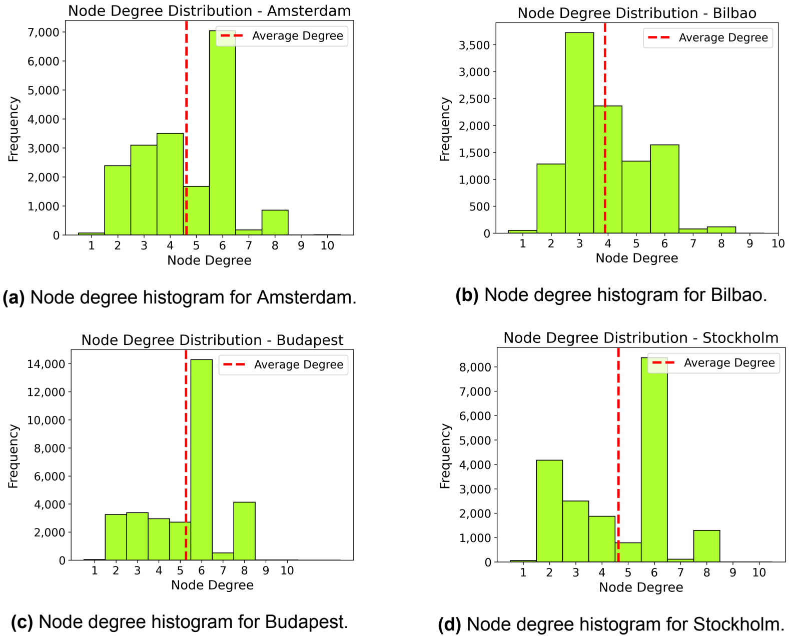
The five road topology indicators capture key aspects of connectivity, efficiency, and robustness within the road networks. Their values for Amsterdam, Bilbao, Budapest, and Stockholm are summarized in Table 4.1.

Table 4.1: Road topology indicators for Amsterdam, Bilbao, Budapest, and Stockholm.

Indicator	Name	Amsterdam	Bilbao	Budapest	Stockholm
k_μ	Mean node degree	4.622	3.895	5.264	4.628
k_{cv}	Coefficient of variation of node degree	0.358	0.350	0.338	0.408
R	Network efficiency ratio	0.430	0.446	0.474	0.422
$C_{B,\mathrm{cv}}$	Coefficient of variation of betweenness centrality	2.821	2.691	3.745	4.492
$C_{B,95}$	95th percentile of betweenness centrality	0.018	0.026	0.010	0.012

Among these cities, Budapest exhibits the highest average node degree (k_μ), suggesting a more interconnected network structure, while Bilbao shows the lowest. Regarding the variability in node degree (k_{cv}), all four cities display relatively similar levels, although Stockholm shows slightly higher variability. The degree distributions, shown in Figure 4.1, illustrate how Amsterdam and Budapest exhibit more uniform connectivity patterns compared to Bilbao and Stockholm.

Network efficiency (R) is highest in Budapest, suggesting that average travel distances are relatively short, while Stockholm requires the largest detours. In terms of variability in betweenness centrality



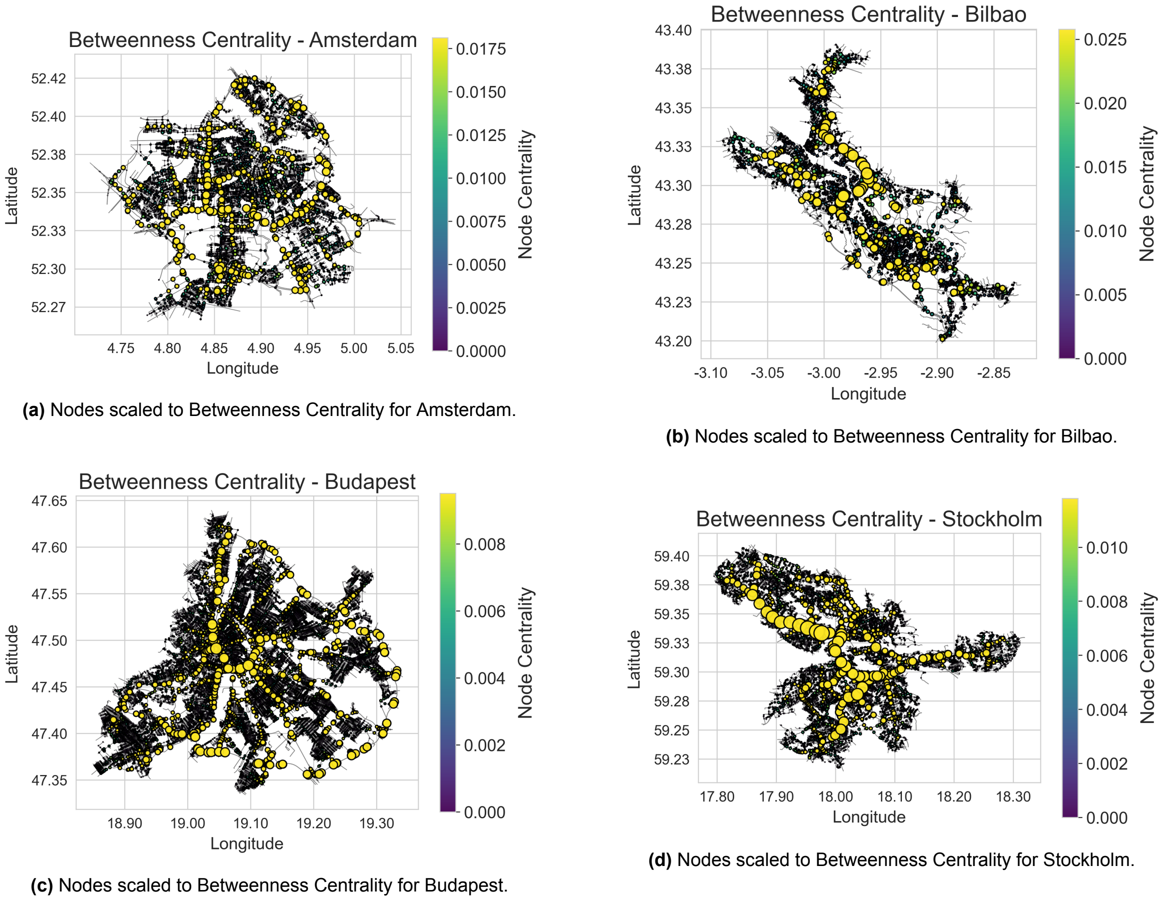
(a) Node degree histogram for Amsterdam.

(b) Node degree histogram for Bilbao.

(c) Node degree histogram for Budapest.

(d) Node degree histogram for Stockholm.

Figure 4.1: Node degree distributions for (a) Amsterdam, (b) Bilbao, (c) Budapest, and (d) Stockholm.



(a) Nodes scaled to Betweenness Centrality for Amsterdam.

(b) Nodes scaled to Betweenness Centrality for Bilbao.

(c) Nodes scaled to Betweenness Centrality for Budapest.

(d) Nodes scaled to Betweenness Centrality for Stockholm.

Figure 4.2: Betweenness Centrality maps for (a) Amsterdam, (b) Bilbao, (c) Budapest, and (d) Stockholm. The colorbar is capped at the 95th percentile.

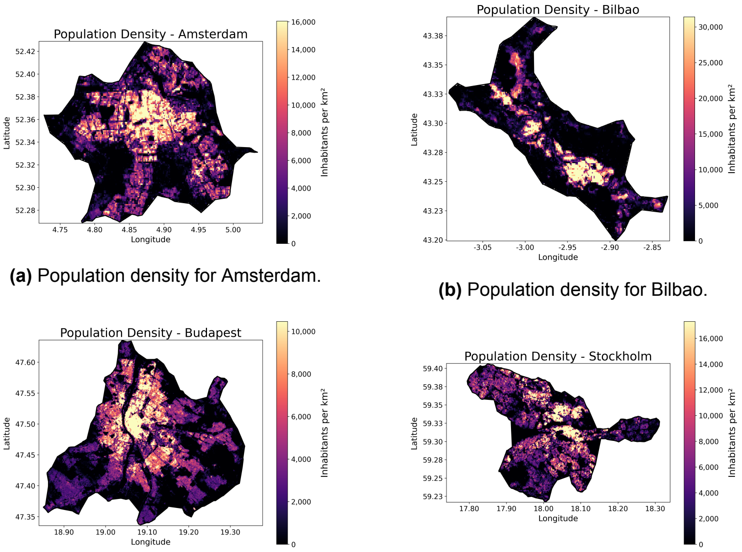
($C_{B,\mathrm{cv}}$), Stockholm exhibits the highest variation among the four cities. This pattern can be attributed to its geographic constraints, where islands and water crossings channel trips through a limited number of key intersections. Amsterdam also shows relatively high variability, although to a lesser extent. In contrast, Budapest and Bilbao display lower variability in betweenness, indicating a more evenly distributed network of important nodes.

The 95th percentile of betweenness centrality ($C_{B,95}$) further highlights differences between the cities. Bilbao stands out with the highest value, indicating a stronger reliance on a few critical intersections for overall connectivity. Stockholm and Amsterdam follow with moderate values, while Budapest records the lowest value, reflecting a more balanced distribution of traffic flows across its network. These spatial patterns are illustrated in Figure 4.2.

The number and spatial distribution of inhabitants and economic activities provide valuable insights into potential movement patterns. Table 4.2 summarizes the indicators related to population, shops, and offices for the four selected cities, highlighting both the volume and spatial variability of activities within the urban areas.

Table 4.2: Population and facility indicators for Amsterdam, Bilbao, Budapest, and Stockholm.

Indicator	Name	Amsterdam	Bilbao	Budapest	Stockholm
P_μ	Mean population density (100×100m grid)	4,365.0	5,601.6	2,859.1	3,167.1
P_{cv}	Coefficient of variation of population density	1.261	1.931	1.481	1.652
P_{95}	95th percentile of population density	16,056.6	31,407.1	10,465.6	14,566.1
S_μ	Mean shop density (500×500m grid)	28.4	32.6	21.0	22.2
S_{cv}	Coefficient of variation of shop density	2.970	3.323	2.730	2.760
O_μ	Mean office density (500×500m grid)	6.5	6.7	6.2	5.9
O_{cv}	Coefficient of variation of office density	1.354	1.747	1.557	1.458



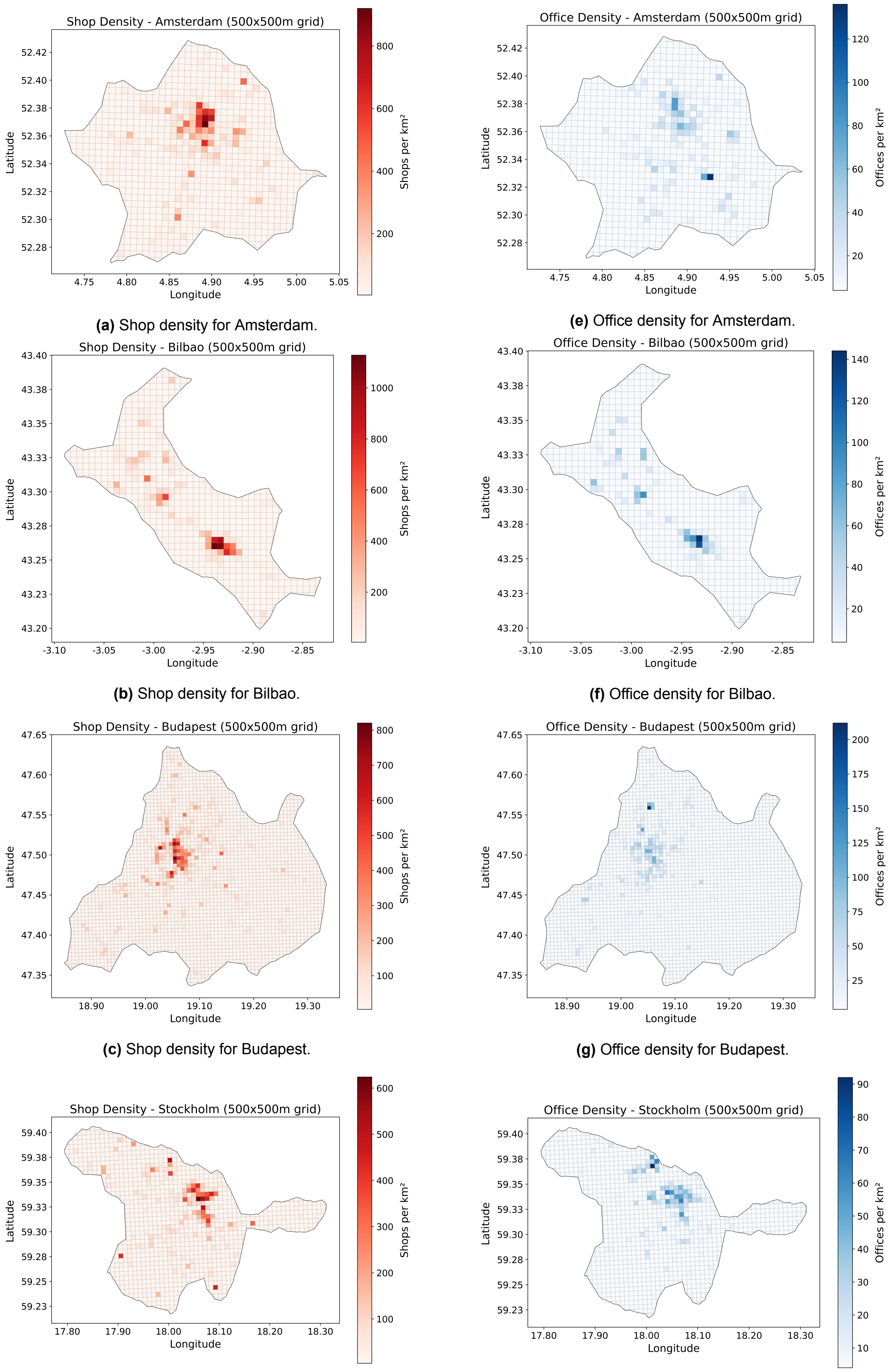
(a) Population density for Amsterdam.

(b) Population density for Bilbao.

(c) Population density for Budapest.

(d) Population density for Stockholm.

Figure 4.3: Population density maps (inhabitants/km²; 100×100m grid) for (a) Amsterdam, (b) Bilbao, (c) Budapest, and (d) Stockholm. The colorbar is capped at the 95th percentile.



(a) Shop density for Amsterdam.

(e) Office density for Amsterdam.

(b) Shop density for Bilbao.

(f) Office density for Bilbao.

(c) Shop density for Budapest.

(g) Office density for Budapest.

(d) Shop density for Stockholm.

(h) Office density for Stockholm.

Figure 4.4: (a-d) Shop and (e-h) office density maps (locations/km²; 500×500m grid) for (a,e) Amsterdam, (b,f) Bilbao, (c,g) Budapest, and (d,h) Stockholm.

Bilbao shows the highest mean population density (P_μ) while it is the smallest of the four cities, suggesting a more concentrated residential structure. Budapest records the lowest, indicating a more dispersed pattern. Amsterdam and Stockholm lie between these two cities. The spatial distribution of population density is illustrated in Figure 4.3, highlighting how residents are spread across the urban areas. The coefficient of variation of population density (P_{cv}) further supports these observations: Bilbao and Stockholm display higher variability, indicating that population is more concentrated in specific zones, whereas Amsterdam and Budapest have lower variability, reflecting a more balanced distribution across their urban area.
The 95th percentile of population density (P_{95}) captures the density levels in the most populated areas. Bilbao again shows the highest value, followed by Stockholm and Amsterdam with moderate levels. Budapest exhibits the lowest value, consistent with its more evenly spread population and lower overall intensity.

A similar pattern emerges for the average shop density (S_μ), with Bilbao and Amsterdam recording the highest levels of commercial activity, followed by Stockholm and Budapest. Despite differences in average values, the coefficient of variation of shop density (S_{cv}) is relatively high for all four cities, indicating that commercial activity tends to be concentrated in specific zones rather than evenly distributed across the urban area.

Office density (O_μ) shows less variation between the cities, with similar average values per km^2. Among the four, Amsterdam records the highest office density, while Stockholm has the lowest. However, the coefficient of variation of office density (O_{cv}) reveals more pronounced differences: Budapest and Bilbao exhibit greater variability, suggesting that offices are clustered into specific business districts, while Amsterdam and Stockholm show lower variability, pointing to a more even spatial distribution.

The spatial patterns of both shops and offices are illustrated in Figure 4.4. Although some overlap between commercial and office concentrations is visible, the clusters are not always aligned. In Amsterdam, for instance, the main office hub is distinct from the primary shopping areas, highlighting functional specialization within the urban areas.

The indicators in Table 4.3 show variability in mobility and congestion across the four example cities. The share of trips made by motorized vehicles (M_{MV}) varies notably, with Stockholm showing the highest reliance on private motorized transport, while Bilbao exhibits the lowest share, suggesting a less car-oriented mobility pattern.

In contrast, the share of trips by public transport (M_{PT}) is highest in Budapest and Stockholm, indicating that public transport plays an important role in daily trips. Amsterdam and Bilbao record lower public transport shares, which may reflect either limited service levels or greater attractiveness of alternative modes.

The share of active modes (M_{AM}), including walking and cycling, shows even sharper contrasts. Amsterdam leads in the use of active modes, likely due to its compact urban form and dedicated infrastructure, while Budapest and Stockholm have relatively low shares. These variations highlight the influence of both infrastructure and cultural factors on urban mobility patterns.

Table 4.3: Mobility and congestion indicators for Amsterdam, Bilbao, Budapest, and Stockholm.

Indicator	Name	Amsterdam	Bilbao	Budapest	Stockholm
M_{MV}	Share of trips with motorized vehicles (%)	33	12	35	39
M_{PT}	Share of trips with public transport (%)	10	22	47	47
M_{AM}	Share of trips with active modes (walking, cycling) (%)	57	66	18	13
V_{own}	Car ownership (cars per 1,000 inhabitants)	632	440	410	267
CL	Congestion level (%)	24	13	32	20

Car ownership (V_{own}), measured as the number of cars per 1,000 inhabitants, also shows variability across the four cities. Amsterdam has the highest ownership rate, despite its high share of active mobility, while Stockholm records the lowest, consistent with its strong reliance on public transport. These differences suggest that car ownership levels do not always align with actual usage patterns and may instead reflect lifestyle preferences or wealth factors. The relationship between mobility profiles and car ownership is shown in Figure 4.5, where the cities are scaled to ownership in a ternary plot. While one might expect a higher number of cars to directly correspond to a higher share of car use, this pattern does not consistently appear across the 32 cities.
Finally, congestion levels (CL) also differ between the four cities. Budapest faces relatively high congestion, indicating possible infrastructure bottlenecks or higher traffic volumes. In contrast, Bilbao and Stockholm report lower congestion levels, suggesting a more efficient distribution of trips across available modes or better traffic management.

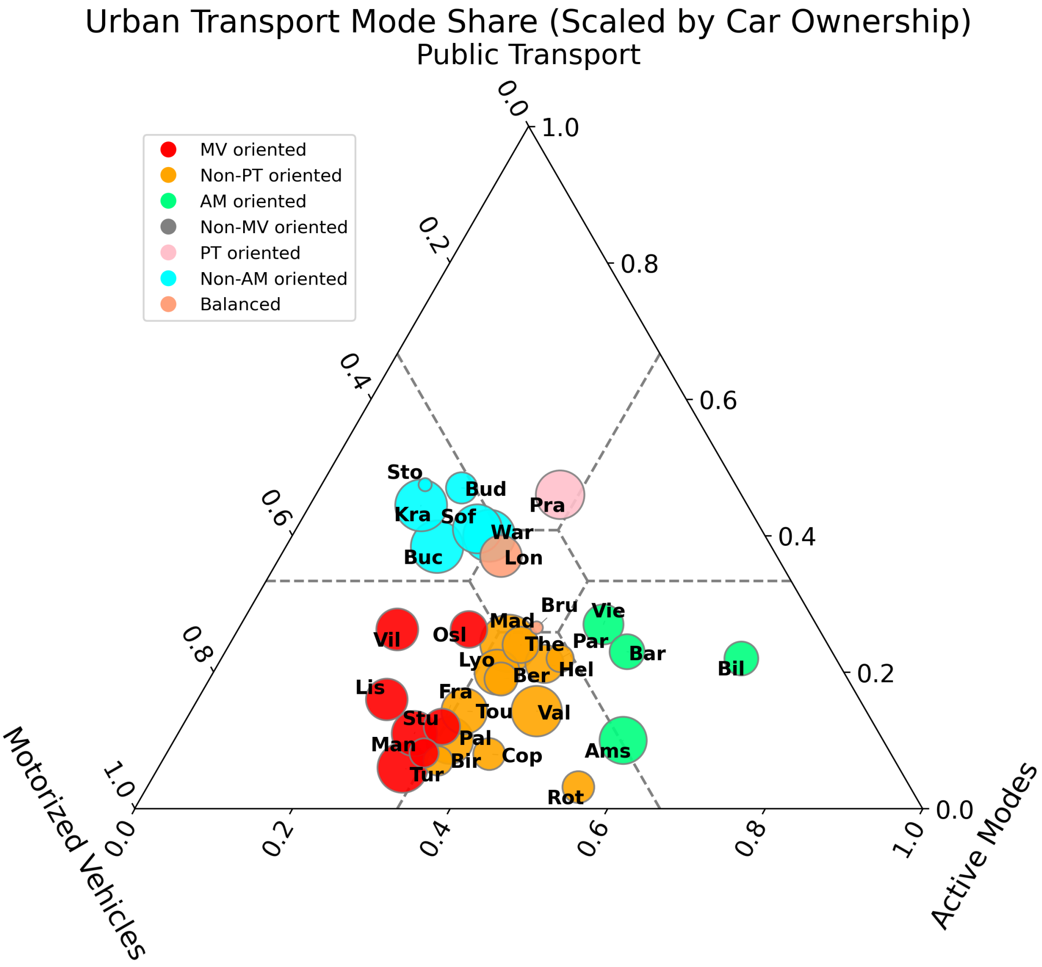


Figure 4.5: Modal share profiles scaled to car ownership across the 32 cities.

When considering all 32 cities, the coefficients of variation in Table 4.4 confirm that the indicators span a wide range of characteristics. The road topology indicators show relatively low variability across cities, including the average node degree (k_μ), its variability (k_{cv}), and the efficiency ratio (R). While this limits their ability to distinguish between cities, these indicators still describe important structural aspects of road networks. In contrast, mobility, population, and economic activity indicators tend to show moderate to high variation. For example, mean population density (P_μ), shop density (S_μ), and office distribution (O_{cv}) vary notably, reflecting the diversity in land-use patterns.

Some indicators stand out with very high variability, including the 95th percentile of betweenness centrality ($C_{B,95}$) and the share of public transport trips (M_{PT}). These indicators have values of the

Table 4.4: Coefficient of variation (CV) for all 17 indicators for the 32 European cities.

Indicator	CV	Indicator	CV	Indicator	CV
k_μ	0.067	P_μ	0.434	M_{MV}	0.273
k_{cv}	0.077	P_{cv}	0.207	M_{PT}	0.568
R	0.070	P_{95}	0.391	M_{AM}	0.342
$C_{B,\mathrm{cv}}$	0.255	S_μ	0.358	V_{own}	0.260
$C_{B,95}$	0.653	S_{cv}	0.222	CL	0.256
		O_μ	0.242		
		O_{cv}	0.359		

coefficient of variation above 0.5, indicating that cities differ substantially in terms of network centrality and public transit reliance. In contrast, car ownership (V_{own}) and congestion level (CL) show more moderate variation. Taken together, the range of CV values demonstrates that the selected indicators are well-suited to capture both subtle and pronounced differences across urban areas. This provides a strong foundation for clustering cities based on their network, population and activity distribution, and mobility characteristics.

4.2. Correlation Analysis

To support effective clustering, it is important to reduce redundancy by identifying and removing certain correlated indicators, which may contain overlapping information and skew the results. As a first step toward the reduction of dimensionality, correlations between all 17 indicators are analyzed using the Pearson correlation coefficient. A coefficient of 1.0 represents a perfect positive correlation (as seen when an indicator is compared with itself), while a value of -1.0 indicates a perfect negative correlation. Many indicator pairs show weak or negligible correlations, suggesting that they capture different aspects of network structure, population and facilities, and mobility. However, several indicators do exhibit stronger correlation relationships. The full correlation matrix can be seen in Appendix B.

To focus on more meaningful relationships, statistically insignificant correlations were filtered out. As explained in Section 3.4, the significance threshold was based on a p-value of 0.05 and a sample size of 32 cities, resulting in a minimum absolute correlation value of 0.349. Correlations below this threshold are not considered meaningful and are excluded from interpretation. The filtered matrix in Figure 4.6 only highlights the significant correlations, which makes identifying strong relations between indicators easier.

Cross-Domain Relationships

Several notable relationships emerge between indicators from different domains, highlighting how road topology, population density, economic activity, and mobility patterns are interconnected.

One example is the positive correlation between congestion levels (CL) and mean node degree (k_μ). While this may seem counterintuitive, it likely reflects that more connected street networks, especially in dense urban cores, attract higher traffic volumes. These complex networks are typically found in areas with high travel demand. This relationship is consistent with Braess's Paradox, which shows that increasing network connectivity does not always improve traffic flow. In some cases, it can even lead to longer travel times due to individually optimal but collectively inefficient route choices (Braess, 1968). The negative correlation between k_μ and the modal share of active modes (M_{AM}) further supports this idea. A greater number of routing options can discourage walking and cycling, possibly due to the dominance of car traffic. Limiting route options for cars in highly connected networks could help ease congestion and encourage more sustainable travel behavior.

Interestingly, the 95th percentile of population density (P_{95}) also correlates negatively with mean node degree (k_μ) and its coefficient of variation (k_{cv}). This indicates that cities with highly concentrated populations tend to have more tree-like or radial street networks, with fewer connections per node, as discussed in Section 2.3. This pattern may reflect historic European urban layouts, where dense central areas are characterized by narrow, organic street patterns rather than uniform grids.

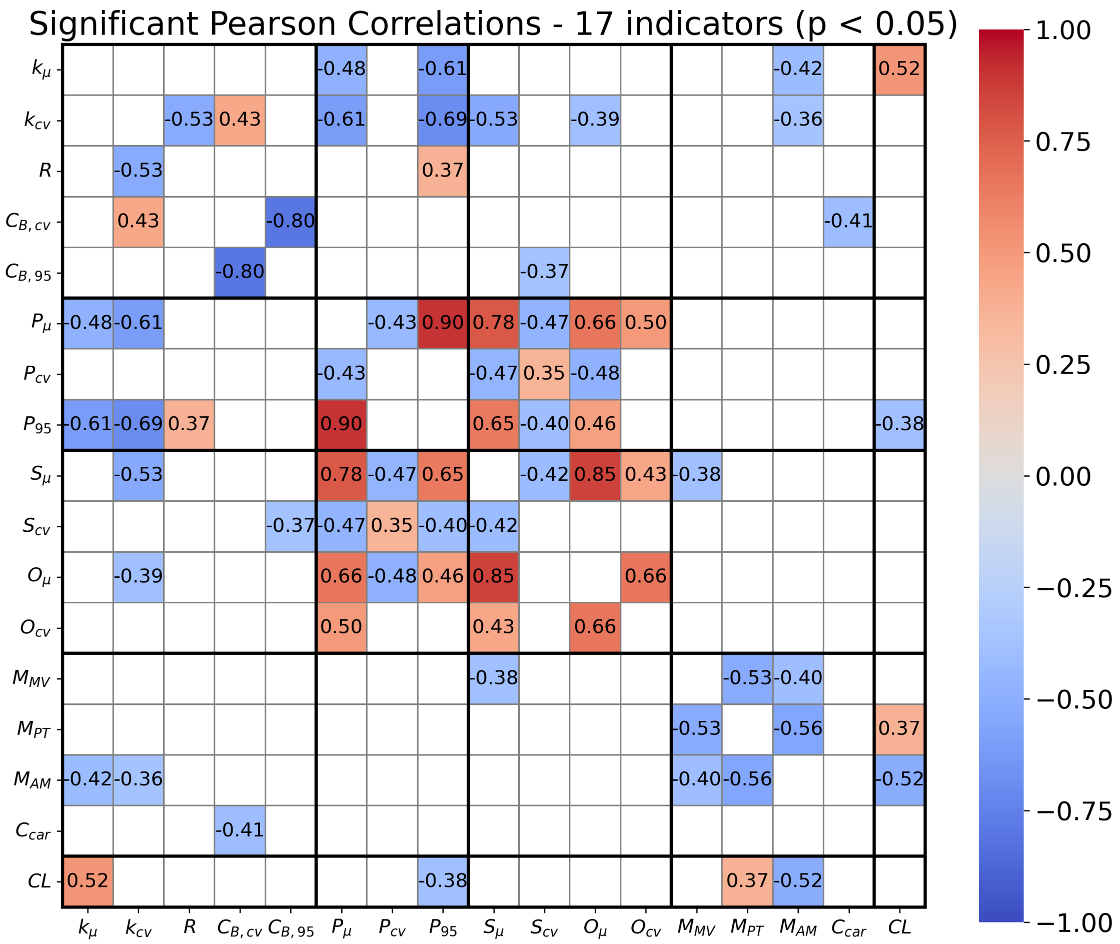


Figure 4.6: Pearson correlation matrix showing only significant correlations between all 17 indicators.

Another clear relationship is observed between mean population density (P_μ) and both shop density (S_μ) and office density (O_μ). This is intuitive, as cities with higher residential densities often concentrate economic activities in close proximity to living areas (Kopczewska et al., 2024; Samburu et al., 2023). Higher densities create greater demand for goods and services, which supports the presence of shop and office facilities in the same urban zones. According to Gehl (2010), this co-location pattern is typical of compact European city centers and reflects the spatial integration of land use.

Overall, these cross-domain correlations confirm that road topology, population distribution, economic activity, and mobility behavior are interconnected within the European urban context.

Indicator Exclusion

Based on the significant correlations, a selection process was applied to reduce redundancy among indicators. Strongly correlated indicators often capture underlying patterns, and thus retaining only one from each correlated group helps to minimize overlap in information while preserving important insights. As a general principle, when strong correlations are observed, preference is given to the indicator showing the highest variability, as shown in Table 4.4, in order to maximize differentiation. Due to potential data limitations related to population, shop and office indicators, discussed in Section 4.1, average values for these indicators are considered less reliable and are thus given lower priority, particularly when they are strongly correlated with more robust indicators. For modal share indicators, only one mode is preserved for the clustering, as the remaining shares can typically be deduced either from that mode or from related indicators such as the congestion level. The following paragraphs outline the specific indicators that were excluded, along with the explanation for their removal, based on their strongest correlations and their role in urban transportation.

$\boldsymbol{k_{\mathrm{cv}}}$ – The coefficient of variation of node degree is strongly negatively correlated with the 95th percentile of population density (P_{95}). Cities with highly variable node degrees often have tree-like or radial street networks, which tend to align with uneven population distributions as discussed in the previous paragraph. Higher population densities are typically associated with more uniform node degrees, as shown by Lin and Ban (2017). A lower k_{cv} often indicates grid-like network structure that supports consistent connectivity and better accessibility. In contrast, networks with more low-degree nodes may not have sufficient route options, possibly requiring detours and reducing overall efficiency of the network. The correlation with the efficiency ratio (R) reinforces this, indicating that more balanced connectivity can facilitate shorter paths. As k_{cv} is well explained by both P_{95} and R, it is excluded.

$\boldsymbol{C_{\mathrm{B,cv}}}$ – The coefficient of variation of betweenness centrality exhibits a strong negative correlation with the 95th percentile of betweenness centrality ($C_{B,95}$). This is expected, as both indicators describe an aspect of the distribution of betweenness centrality within the network. Cities with a few dominant intersections naturally display lower variability in the rest of the network. In other words, if a small number of nodes carry a large share of movements, most other nodes will have low betweenness values, reducing the variability. Furthermore, the negative correlation with car ownership (V_{own}) may reflect the fact that cities with more centralized movement patterns tend to discourage car use due to bottlenecks, or possibly intentional network design. As $C_{B,95}$ captures the skew in network centrality more directly and has high variability, $C_{B,\mathrm{cv}}$ is excluded.

$\boldsymbol{P_{\mu}}$ – The mean population density is very strongly correlated with its 95th percentile (P_{95}), indicating that cities with high average density also tend to have highly concentrated population centers. This redundancy justifies excluding the mean in favor of the percentile, which better captures the presence of dense urban centers. In addition, P_{μ} is negatively correlated with the coefficient of variation of population (P_{cv}), indicating that within the 32 cities with higher average density have more evenly distributed populations.

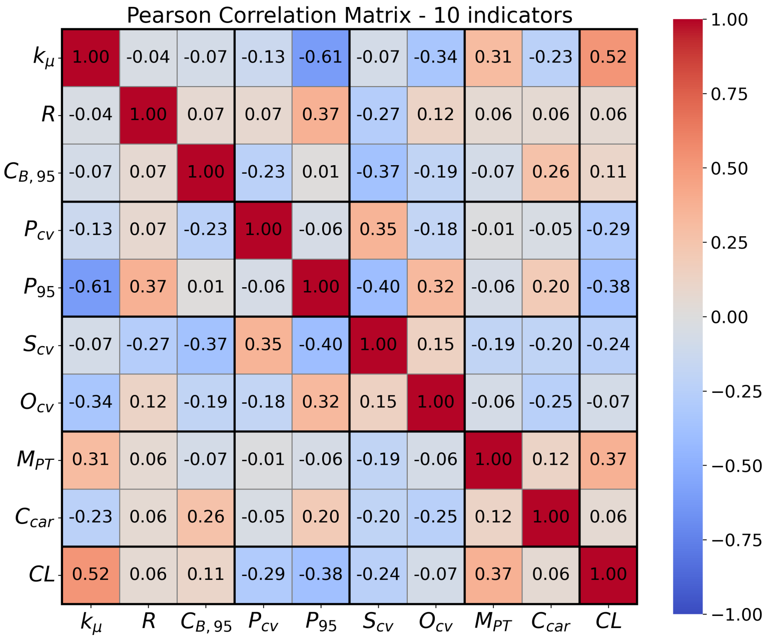
$\boldsymbol{S_{\mu}}$ – The average density of shops is strongly positively correlated with P_{95}, reflecting that retail activity tends to cluster in densely populated urban areas (Samburu et al., 2023). It also correlates with the coefficients of variation for offices and population, highlighting the spatial overlap between residential and commercial functions. Its negative correlation with S_{cv} indicates that higher average shop density typically corresponds with a more even distribution of retail facilities. Since S_{cv} captures this dispersion, and P_{95} reflects the spatial demand, S_{μ} is excluded.

$\boldsymbol{O_{\mu}}$ – Mean office density correlates positively with both the variation in office distribution (O_{cv}) and with population concentration (P_{95}), consistent with patterns of employment clustering in dense urban cores. Companies often locate offices in the same areas to benefit from shared infrastructure and access to the local workforce (Kopczewska et al., 2024). The moderate negative correlation with P_{cv} further implies that cities with clustered employment tend to have more balanced population distributions. Given these overlaps, O_{μ} is excluded.

$\boldsymbol{M_{\mathrm{MV}}}$ – The modal share of motorized vehicles is logically correlated with the other two modal shares, as the three percentages sum to 100%. A decrease in one mode necessarily results in an increase in one or both of the others. As a result, it is sufficient to retain only one modal share to capture most of the variation. In this case, M_{PT} is retained due to its stronger correlations with the other two modal share indicators, and its higher coefficient of variation. Thus, M_{MV} is excluded from further analysis.

$\boldsymbol{M_{\mathrm{AM}}}$ – The modal share for active modes is also excluded, not only due to its strong correlation with M_{PT}, but also because it exhibits a notable negative correlation with the congestion level (CL). This relationship is consistent with theoretical expectations, as higher shares of walking and cycling tend to reduce reliance on motorized transport, thereby reducing pressure on road networks. Such dynamics have been observed in the literature, as highlighted by Rabl and De Nazelle (2012), where increased active mobility is associated with lower congestion impacts.

With the seven indicators excluded, the other ten remain for further analysis. The correlation matrix in Figure 4.7 shows that some significant correlations still persist, indicating potential redundancy. To remove this remaining overlap and extract independent patterns, a Principal Component Analysis (PCA) is performed in Section 4.3.



	k_μ	R	$C_{B,95}$	P_{cv}	P_{95}	S_{cv}	O_{cv}	M_{PT}	C_{car}	CL
k_μ	1.00	-0.04	-0.07	-0.13	-0.61	-0.07	-0.34	0.31	-0.23	0.52
R	-0.04	1.00	0.07	0.07	0.37	-0.27	0.12	0.06	0.06	0.06
$C_{B,95}$	-0.07	0.07	1.00	-0.23	0.01	-0.37	-0.19	-0.07	0.26	0.11
P_{cv}	-0.13	0.07	-0.23	1.00	-0.06	0.35	-0.18	-0.01	-0.05	-0.29
P_{95}	-0.61	0.37	0.01	-0.06	1.00	-0.40	0.32	-0.06	0.20	-0.38
S_{cv}	-0.07	-0.27	-0.37	0.35	-0.40	1.00	0.15	-0.19	-0.20	-0.24
O_{cv}	-0.34	0.12	-0.19	-0.18	0.32	0.15	1.00	-0.06	-0.25	-0.07
M_{PT}	0.31	0.06	-0.07	-0.01	-0.06	-0.19	-0.06	1.00	0.12	0.37
C_{car}	-0.23	0.06	0.26	-0.05	0.20	-0.20	-0.25	0.12	1.00	0.06
CL	0.52	0.06	0.11	-0.29	-0.38	-0.24	-0.07	0.37	0.06	1.00

Figure 4.7: Pearson correlation matrix for the 10 included indicators.

4.3. Principal Component Analysis

To remove redundancy between indicators and reveal distinct patterns for clustering, a Principal Component Analysis (PCA) was performed. As described in Section 3.4, PCA transforms the original set of potentially correlated indicators into a set of principal components (PCs). Each principal component is uncorrelated with the other components and captures a distinct portion of the dataset's variance, effectively eliminating redundancy while preserving the underlying information. This dimensionality reduction avoids overlap between indicators and ensures that clustering is based on independent patterns in the data.

PCA was applied to the included set of ten indicators. The resulting scree plot in Figure 4.8 illustrates the proportion of variance explained by each principal component, along with the cumulative variance across all components. A threshold of 75% cumulative explained variance was set as the minimum required to justify dimensionality reduction for the clustering analysis. Based on this criterion, the first five principal components were selected. Together, they account for 79.3% of the total variance, indicating that the reduced-dimensional representation preserves most of the original variation in the dataset.

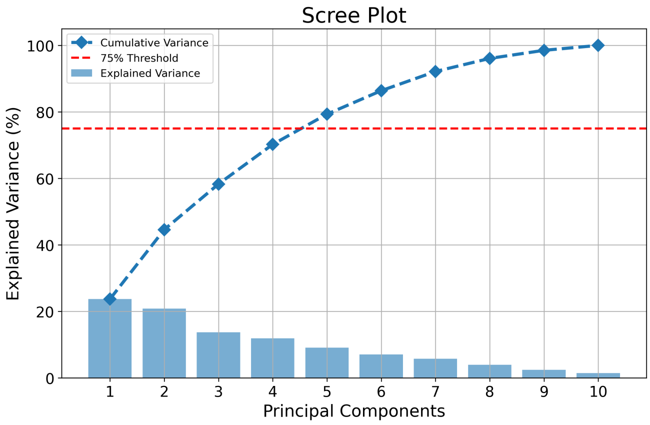
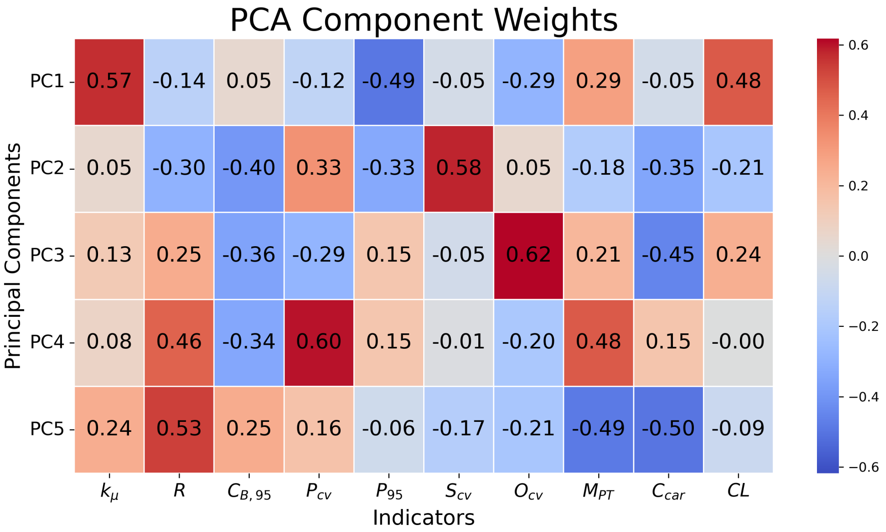


Figure 4.8: Scree plot showing the explained variance per principal component, and the cumulative explained variance. The dashed red line indicates the minimal necessary explained variance for representative results.

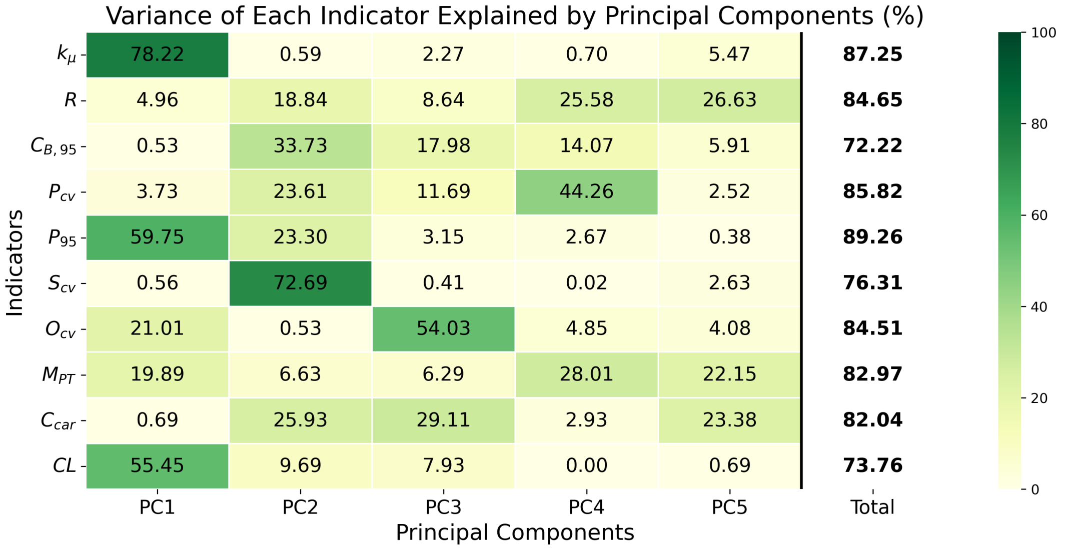
To understand how the five retained principal components relate to the original ten indicators, the loadings of the indicators are presented in Figure 4.9, which reflect the strength and direction of the relationships between each indicator and principal component. Positive loading values indicate a positive relationship, while negative values indicate a negative relationship. Although they are not direct correlation coefficients, the loadings show how much each original indicator contributes to the definition and interpretation of a principal component. For example, PC1 is shaped by strong positive loadings from k_μ and CL, while P_{95} contributes negatively. PC3 is primarily influenced by O_{cv}, and PC4 shows strong positive associations with both P_{cv} and M_{PT}. This overview helps identify which indicators dominate the construction of each principal component.



	k_μ	R	$C_{B,95}$	P_{cv}	P_{95}	S_{cv}	O_{cv}	M_{PT}	C_{car}	CL
PC1	0.57	-0.14	0.05	-0.12	-0.49	-0.05	-0.29	0.29	-0.05	0.48
PC2	0.05	-0.30	-0.40	0.33	-0.33	0.58	0.05	-0.18	-0.35	-0.21
PC3	0.13	0.25	-0.36	-0.29	0.15	-0.05	0.62	0.21	-0.45	0.24
PC4	0.08	0.46	-0.34	0.60	0.15	-0.01	-0.20	0.48	0.15	-0.00
PC5	0.24	0.53	0.25	0.16	-0.06	-0.17	-0.21	-0.49	-0.50	-0.09

Figure 4.9: Weight factors per indicator for each included principal component.

An alternative way to assess how the five principal components relate to the original indicators is to examine the share of each indicator's variance explained by each component. Figure 4.10 presents this information as a percentage per principal component, along with the total variance explained for all included indicators. This perspective complements the interpretation of loadings by showing not just how indicators shape the components, but also how well the variation of each indicator is represented in the reduced-dimensional space.



	PC1	PC2	PC3	PC4	PC5	Total
k_μ	78.22	0.59	2.27	0.70	5.47	**87.25**
R	4.96	18.84	8.64	25.58	26.63	**84.65**
$C_{B,95}$	0.53	33.73	17.98	14.07	5.91	**72.22**
P_{cv}	3.73	23.61	11.69	44.26	2.52	**85.82**
P_{95}	59.75	23.30	3.15	2.67	0.38	**89.26**
S_{cv}	0.56	72.69	0.41	0.02	2.63	**76.31**
O_{cv}	21.01	0.53	54.03	4.85	4.08	**84.51**
M_{PT}	19.89	6.63	6.29	28.01	22.15	**82.97**
C_{car}	0.69	25.93	29.11	2.93	23.38	**82.04**
CL	55.45	9.69	7.93	0.00	0.69	**73.76**

Figure 4.10: Variance captured per indicator by each included principal component.

While some indicators are strongly associated with a single principal component, such as k_μ with PC1 and S_{cv} with PC2, others show a more distributed pattern of explained variance. For instance, indicators like R, B_{95}, M_{PT}, and V_{own} are each partially captured by several components, making their interpretation more complex. Despite this, all indicators have more than 70% of their variance explained by the selected principal components, confirming that the dimensionality reduction retains a high level of information for all ten indicators. It is worth noting that CL, although strongly represented in PC1, has one of the lowest total explained variances across all components, just slightly higher than B_{95}, which is the least well-represented overall (72.22%). In the next paragraphs, each principal component is elaborated on.

PC1 – This principal component accounts for a substantial share of the variance in k_μ, P_{95}, and CL, indicating that it reflects aspects of connectivity, population hubs, and congestion. It also explains a moderate share of the variance for O_{cv} and M_{PT}, emphasizing the wide reach of the first component.

PC2 – Dominated by a strong contribution of S_{cv}, this component focuses on shop distribution. It also captures relevant shares of variance for B_{95}, V_{own}, P_{cv}, P_{95}, and R, highlighting its wider relevance across indicators.

PC3 – This component primarily reflects variation in O_{cv}, highlighting the spatial distribution of offices. It also captures smaller but notable shares of variance in V_{own} and B_{95}.

PC4 – A large share of the variance in P_{cv} is explained by this component, along with roughly a quarter of the variance in both R and M_{PT}. PC4 is thus important for capturing differences in population distribution and public transport use.

PC5 – While this component explains a smaller portion of the overall variance, it still adds value by capturing remaining variation in R, M_{PT}, and V_{own}. It accounts for residual patterns not fully covered by the earlier components.

The selected principal components provide a robust and compact representation of the dataset, capturing the most relevant variation across cities while reducing dimensionality and removing redundancy. This reduced structure forms a strong foundation for the clustering analysis in Chapter 5, enabling a clearer analysis of similarities and differences among the included urban areas.

5 Clustering Results

Following the principal component analysis in Chapter 4, this chapter presents the clustering results for the 32 cities based on their road topology, activity distribution, and mobility characteristics. Three clustering approaches were applied: K-Means, K-Medoids, and hierarchical clustering using Ward's method. Based on the evaluation of elbow plots, silhouette scores, and dendrogram structure, configurations with two, five, and seven clusters emerged as the most meaningful. The selection of the optimal number of clusters is shown in Section 5.1.
After determining the optimal number of clusters, particular attention is given to the seven-cluster result in Section 5.2, which is analyzed in detail based on the full agreement across methods. The two- and five-cluster results are subsequently discussed, offering complementary perspectives.

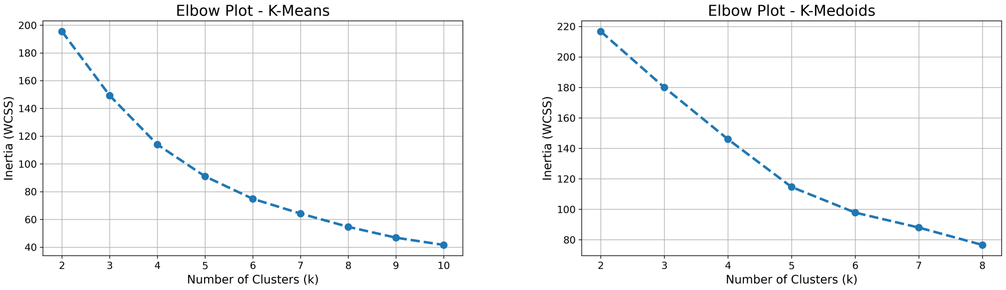
5.1. Number of Clusters

To determine the most suitable number of clusters, the clustering outcomes were evaluated using a combination of elbow plots, silhouette scores, and dendrogram structure. These methods provide insights into the internal coherence and separation between the clusters. Based on these evaluations, the configurations with two, five, and seven clusters were identified as the most promising. First, the results for the K-clustering methods, K-Means and K-Medoids, are presented in subsection 5.1.1, where a comparison between the methods is also discussed. The hierarchical clustering results based on Ward's method are elaborated in subsection 5.1.2.

5.1.1. K-Means & K-Medoids

The K-Means and K-Medoids clustering methods are first evaluated. While both approaches divide the dataset into k groups based on similarity, they differ in that K-Means allows cluster centers to lie outside the set of observations, whereas K-Medoids restricts centers to actual data points (medoids). As discussed in Section 3.4, this distinction can lead to differences in cluster compactness and robustness. K-Means clustering was evaluated across a range of $k = 2$ to $k = 10$ clusters, while for K-Medoids, values beyond $k = 8$ were not considered due to computational limitations. This restriction is justified by the K-Means results, which suggests that meaningful cluster configurations are captured within this range.

The elbow plots for K-Means and K-Medoids are presented in Figure 5.1. In both cases, the inertia decreases rapidly for low values of k and then begins to level off, forming an elbow-shaped curve that indicates declining improvements in clustering performance as the number of clusters increases. For the K-Means method, a notable inflection point appears around five or six clusters, suggesting that the internal structure of the dataset is well captured with a relatively small number of groups. In contrast, K-Medoids shows consistently higher inertia values, as medoids must be actual data points rather than optimal mathematical centers. This constraint can lead to higher distances to the cluster center, explaining the higher inertia levels observed. The elbow shape is more pronounced for K-Medoids, with a clear inflection at five clusters. Despite the minor differences, both methods indicate that five clusters represent a strong candidate for the structure underlying the dataset.

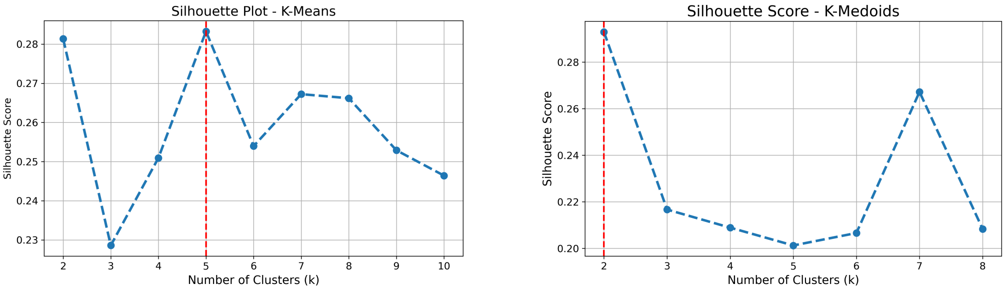


(a) Elbow plot for K-Means.

(b) Elbow plot for K-Medoids.

Figure 5.1: Elbow plots for (a) K-Means and (b) K-Medoids clustering for increasing values of k.

The silhouette scores for K-Means and K-Medoids are shown in Figure 5.2. For K-Means, the highest average silhouette score occurs at five clusters, closely followed by two clusters. The silhouette scores for seven and eight clusters follow in third and fourth position, but their corresponding silhouette scores are notably lower than for two and five clusters. The silhouette scores for K-Medoids show a different pattern: the highest value is found at two clusters, followed by a secondary peak at seven clusters. Interestingly, the silhouette score at five clusters is the lowest across all included k values for K-Medoids, despite the elbow plot suggesting five clusters as a promising configuration. These findings highlight that while both methods point towards two, five, and seven clusters as meaningful options, the relative strength of the five-cluster result differs between K-Means and K-Medoids.



(a) Silhouette scores for K-Means.

(b) Silhouette scores for K-Medoids.

Figure 5.2: Silhouette scores for (a) K-Means and (b) K-Medoids clustering for increasing values of k.

5.1.2. Ward's Method

Hierarchical clustering using Ward’s method was also applied, with evaluation based on silhouette scores for $k = 2$ to $k = 10$ and the resulting dendrogram structure. As shown in Figure 5.3, the silhouette score peaks at seven clusters, closely followed by five and eight, with a notable local maximum at two clusters which aligns with the patterns from K-Means and K-Medoids. These results indicate that five, seven, and eight clusters are most promising.

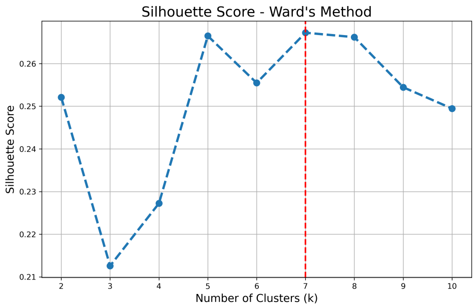


Figure 5.3: Silhouette scores for Ward’s Method for increasing values of k.

To further explore the clustering structure, the complete dendrogram obtained from Ward's method is shown in Figure 5.4. Several distinct divisions are visible at higher linkage distances, indicated by wider horizontal lines. These divisions represent stages where the dissimilarity between merged groups increases substantially, suggesting logical separations between clusters. A clear separation into five and seven clusters can be observed, corresponding to the peaks identified in the silhouette analysis. At eight clusters, however, Budapest separates into its own cluster, providing limited additional insights. Based on these patterns, the five- and seven-cluster configurations are considered the most informative, while the eight-cluster result is excluded from further analysis.

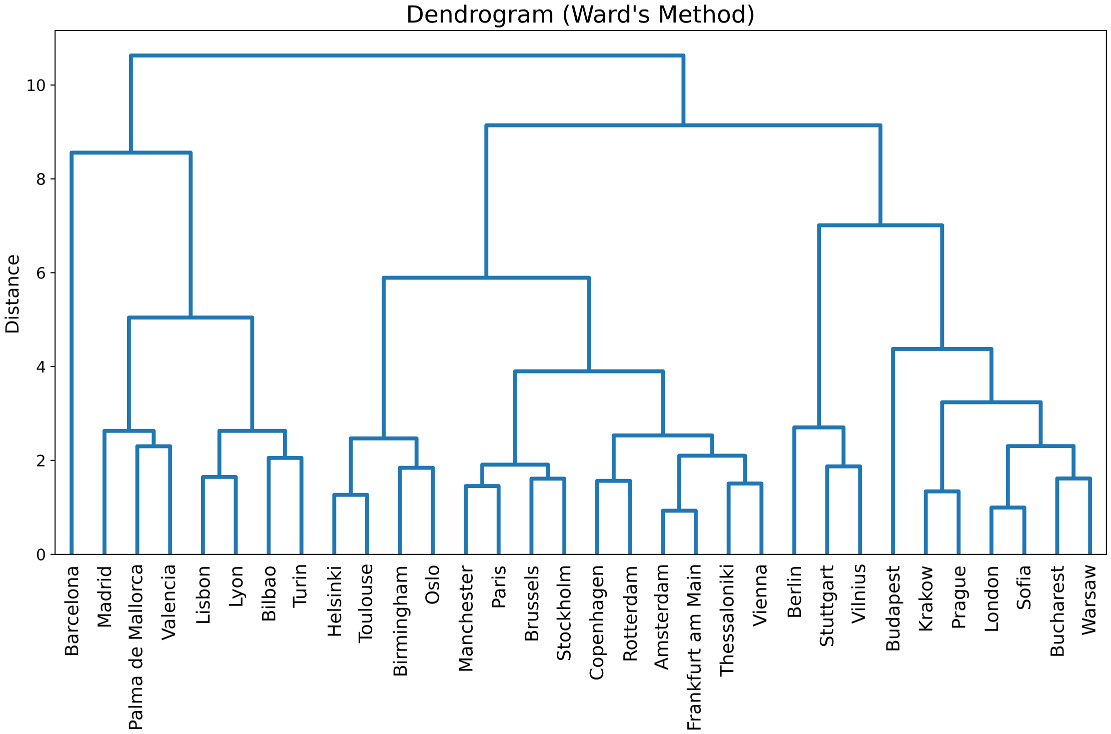
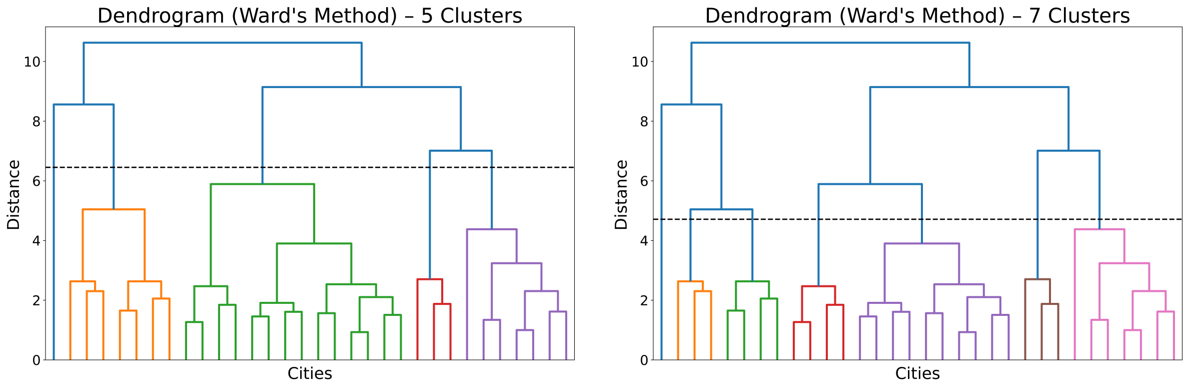


Figure 5.4: Hierarchical clustering dendrogram of the 32 cities using Ward's Method.

To illustrate the clustering structures at specific values of k, Figure 5.5 presents the dendrograms cut at five and seven clusters. These visualizations highlight how the main groups identified through Ward's method evolve when moving from a wider to a more detailed classification.



(a) Dendrogram cut at five clusters.

(b) Dendrogram cut at seven clusters.

Figure 5.5: Ward's Method dendrogram showing the structure at (a) five and (b) seven clusters.

In summary, the clustering evaluation across K-Means, K-Medoids, and Ward's method consistently highlights two, five, and seven clusters as meaningful configurations. Among these, the seven-cluster result proves to be the most robust and informative configuration across methods. Section 5.2 therefore focuses first on analyzing the seven-cluster result in detail, before examining the complementary insights offered by the two- and five-cluster outcomes.

5.2. Cluster Comparison

After identifying two, five, and seven clusters as promising results, this section analyzes the clustering outcomes in greater detail. Given its full cross-method agreement and internal consistency, the seven-cluster result is first analyzed in subsection 5.2.1. Each group is characterized based on road topology, activity distribution, and mobility indicators. Subsequently, the two- and five-cluster results are discussed in subsection 5.2.2 and subsection 5.2.3, providing broader perspectives and highlighting the stability of the clustering outcomes.

5.2.1. Seven-Cluster Result

The clustering results across K-Means, K-Medoids, and Ward's method show perfect agreement for the seven-cluster result, with an Adjusted Rand Index (ARI) of 1.00 between all pairwise comparisons, as shown in the ARI heatmap in Figure 5.6. This result confirms the robustness of the seven-cluster configuration, aligning with the findings from the silhouette score analysis discussed in Section 5.1. Overall, the seven clusters capture the diversity in road topology, activity distribution, and mobility patterns across the 32 cities. The average values and standard deviations for all seventeen indicators across the clusters are presented in Table 5.1. In addition, the seven groups will be described and elaborated on in the following paragraphs.

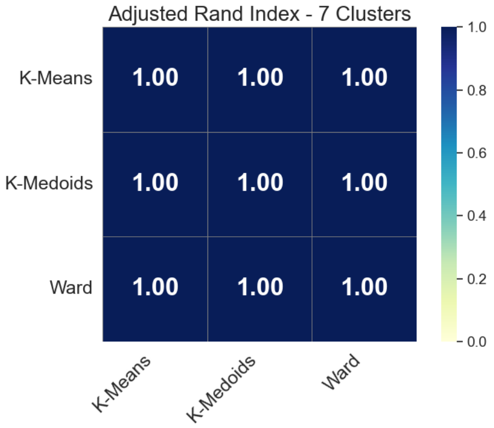


Figure 5.6: Adjusted Rand Index comparing the seven-cluster results obtained from K-Means, K-Medoids, and Ward's method.

Centralized Car-Oriented

4 cities: *Bilbao, Lisbon, Lyon, Turin*

The cities in this cluster are defined by a highly centralized urban form with distinct spatial contrasts. Population distribution is strongly uneven ($P_{\text{cv}} = 1.806$), combining dense cores ($P_{95} = 24,567$ inhabitants/km²) with low-density surrounding areas. Despite their centralization, they remain car-oriented, with a low public transport modal share ($M_{\text{PT}} = 16.0\%$) and relatively high car ownership levels ($V_{\text{own}} = 566.0$ vehicles/1,000 inhabitants). The road networks show moderate connectivity ($k_{\mu} = 4.197$) and average efficiency ($R = 0.433$), while the concentration of traffic flows remains low ($C_{B,95} = 0.016$), suggesting a more distributed use of the network. Interestingly, congestion levels ($CL = 20.2\%$) are lower than might be expected given the car dependency, which may reflect effective traffic management or the availability of multiple routing options within the network.

Table 5.1: Mean values and standard deviations (in parentheses) for all 17 indicators across the seven clusters. Bold indicators were included in the principal component analysis.

Indicator	**Description**	Centralized Car-Oriented	Homogeneous Car-Oriented	Dense Multimodal	PT-Oriented Congested	Well-Connected Bottlenecked	Low-Density Concentrated	Balanced Multimodal
	Number of cities	**4**	**3**	**1**	**7**	**3**	**4**	**10**
$\boldsymbol{k_{\mu}}$	Mean node degree	**4.197 (0.204)**	**3.928 (0.148)**	**3.760**	**4.704 (0.289)**	**4.741 (0.344)**	**4.373 (0.101)**	**4.540 (0.166)**
k_{cv}	Coefficient of variation of node degree	*0.365 (0.015)*	*0.327 (0.014)*	*0.301*	*0.375 (0.024)*	*0.367 (0.032)*	*0.399 (0.030)*	*0.372 (0.029)*
$\boldsymbol{R}$	Network efficiency ratio	**0.433 (0.010)**	**0.443 (0.030)**	**0.471**	**0.434 (0.036)**	**0.427 (0.020)**	**0.377 (0.016)**	**0.433 (0.020)**
$C_{B,\mathrm{cv}}$	Coefficient of variation of betweenness centrality	*3.226 (0.591)*	*2.814 (0.346)*	*2.723*	*3.384 (0.798)*	*2.104 (0.493)*	*3.802 (0.916)*	*3.590 (0.878)*
$\boldsymbol{C_{B,95}}$	95th percentile of betweenness centrality	**0.016 (0.007)**	**0.023 (0.011)**	**0.016**	**0.017 (0.008)**	**0.046 (0.019)**	**0.016 (0.007)**	**0.015 (0.007)**
P_{μ}	Mean population density (100x100m grid)	*4,806.8 (788.0)*	*9,911.7 (3,802.1)*	*20,744.7*	*4,732.4 (1,757.9)*	*6,546.5 (1,191.5)*	*3,249.0 (733.5)*	*5,010.1 (1,246.7)*
$\boldsymbol{P_{\mathrm{cv}}}$	Coefficient of variation of population density	**1.806 (0.216)**	**1.161 (0.233)**	**0.791**	**1.266 (0.148)**	**0.931 (0.101)**	**1.196 (0.243)**	**1.268 (0.146)**
$\boldsymbol{P_{95}}$	95th percentile of population density	**24,567.0 (6,448.7)**	**32,231.5 (6,758.9)**	**49,940.9**	**16,459.5 (5,241.0)**	**18,405.7 (2,694.0)**	**10,400.2 (1,120.8)**	**17,921.6 (4,861.6)**
S_{μ}	Mean shop density (500x500m grid)	*25.1 (5.938)*	*32.1 (14.778)*	*76.6*	*30.9 (9.377)*	*40.8 (14.054)*	*17.1 (2.757)*	*26.8 (9.474)*
$\boldsymbol{S_{\mathrm{cv}}}$	Coefficient of variation of shop density	**3.041 (0.678)**	**2.363 (0.246)**	**1.849**	**2.471 (0.332)**	**1.826 (0.490)**	**3.505 (0.852)**	**2.845 (0.281)**
O_{μ}	Mean office density (500x500m grid)	*7.1 (1.646)*	*7.0 (1.837)*	*19.9*	*7.2 (1.413)*	*12.2 (3.689)*	*7.0 (1.543)*	*7.1 (1.716)*
$\boldsymbol{O_{\mathrm{cv}}}$	Coefficient of variation of office density	**1.660 (0.486)**	**1.269 (0.256)**	**6.251**	**1.525 (0.306)**	**1.460 (0.099)**	**2.386 (1.144)**	**1.666 (0.544)**
M_{MV}	Share of trips with motorized vehicles (%)	*44.8 (23.372)*	*45.3 (8.386)*	*26.0*	*35.3 (6.184)*	*52.0 (7.550)*	*47.2 (8.921)*	*41.2 (10.141)*
$\boldsymbol{M_{\mathrm{PT}}}$	Share of trips with public transport (%)	**16.0 (7.118)**	**16.0 (7.211)**	**23.0**	**41.9 (3.891)**	**18.7 (7.506)**	**17.0 (8.287)**	**18.7 (13.158)**
M_{AM}	Share of trips with active modes (%)	*39.2 (18.500)*	*37.7 (4.619)*	*51.0*	*22.6 (5.968)*	*29.0 (8.544)*	*34.8 (4.924)*	*39.1 (12.161)*
$\boldsymbol{V_{\mathrm{own}}}$	Car ownership (cars per 1,000 inhabitants)	**566.0 (101.390)**	**711.7 (96.345)**	**453.0**	**632.4 (116.151)**	**526.3 (86.558)**	**493.2 (89.749)**	**421.4 (111.417)**
$\boldsymbol{CL}$	Congestion level (%)	**20.2 (4.856)**	**18.7 (0.577)**	**22.0**	**35.0 (5.164)**	**32.0 (5.196)**	**25.8 (5.123)**	**27.1 (4.433)**

Homogeneous Car-Oriented

3 cities: *Madrid, Palma de Mallorca, Valencia*

This cluster groups cities with a relatively homogeneous and compact urban structure. Population densities reach the highest peaks among the multi-city clusters ($P_{95} = 32,231.5$ inhabitants/km²), yet spatial variation is limited ($P_{\text{cv}} = 1.161$), reflecting a dense distribution without strong internal contrasts. Their road networks show the lowest connectivity across all multi-city clusters ($k_{\mu} = 3.928$) but maintain moderate overall efficiency ($R = 0.443$), suggesting that traffic flow can remain effective even though there are fewer route options. Economic activities are evenly distributed, with relatively low variation in both shop ($S_{\text{cv}} = 2.363$) and office ($O_{\text{cv}} = 1.269$) distributions. Mobility patterns reveal a strong dependence on cars: public transport modal share is low ($M_{\text{PT}} = 16.0\%$), and car ownership is the highest recorded among all clusters ($V_{\text{own}} = 711.7$ vehicles/1,000 inhabitants). Despite this car dependence, congestion remains remarkably low ($CL = 18.7\%$), indicating that compactness and potentially efficient traffic management mitigate traffic pressures. The relatively low concentration of traffic ($C_{B,95} = 0.023$) further supports the idea of a functional road network.

Dense Multimodal

1 city: *Barcelona*

This cluster, consisting solely of Barcelona, is defined by extremely high and evenly distributed population densities ($P_{95} = 49,940.9$ inhabitants/km², $P_{\text{cv}} = 0.791$), paired with a compact and efficient road network. While connectivity is low ($k_{\mu} = 3.760$), overall efficiency reaches the highest value among all clusters ($R = 0.471$), supporting direct travel despite the dense environment. Shops are evenly spread ($S_{\text{cv}} = 1.849$), whereas offices are highly concentrated, indicating extreme centralization. Mobility patterns reveal a strong multimodal character: car ownership is low ($V_{\text{own}} = 453.0$ vehicles/1,000 inhabitants), active modes account for more than half of all trips ($M_{\text{AM}} = 51.0\%$), and public transport usage is moderate ($M_{\text{PT}} = 23.0\%$). Despite the extreme densities, congestion ($CL = 22.0\%$) and traffic concentration ($C_{B,95} = 0.016$) remain moderate, suggesting that Barcelona's urban form successfully supports efficient and diverse mobility options.

PT-Oriented Congested

7 cities: *Bucharest, Budapest, Krakow, London, Prague, Sofia, Warsaw*

This cluster brings together Eastern European cities and London, which show strong network connectivity ($k_{\mu} = 4.704$) and moderate efficiency ($R = 0.434$), indicating that while many route options exist, travel paths are not particularly direct. Population densities are moderate ($P_{95} = 16,459.5$ inhabitants/km²) but spatial variation remains relatively high ($P_{\text{cv}} = 1.266$), reflecting clear contrasts between denser centers and sparse suburban neighborhoods. Shops and offices are moderately concentrated ($S_{\text{cv}} = 2.471$, $O_{\text{cv}} = 1.525$). A defining feature of these cities is their strong reliance on public transport: the modal share for public transport ($M_{\text{PT}} = 41.9\%$) is the highest among all clusters. Nevertheless, car ownership is also relatively high ($V_{\text{own}} = 632.4$ vehicles/1,000 inhabitants), and congestion reaches the highest level observed ($CL = 35.0\%$) among all clusters. The concentration of traffic ($C_{B,95} = 0.017$) remains low, indicating that traffic demand is widespread across the network rather than focused on a few intersections, but the overall intensity of travel demand leads to substantial congestion despite the strong role of public transport.

Well-Connected Bottlenecked

3 cities: *Berlin, Stuttgart, Vilnius*

The cities in this cluster have well-connected road networks ($k_{\mu} = 4.741$) but moderate efficiency ($R = 0.427$), indicating that trips often require small detours despite the large number of connections. Traffic flows are notably concentrated, with the most extreme value across all clusters ($C_{B,95} = 0.046$), suggesting that a limited number of intersections carry a disproportionate share of movements. Population density peaks are moderate ($P_{95} = 18,405.7$ inhabitants/km²) and spatial variation is relatively low ($P_{\text{cv}} = 0.931$), pointing to a continuous and evenly developed urban structure. Economic activities are evenly distributed, with little spatial variation for both shops ($S_{\text{cv}} = 1.826$) and offices ($O_{\text{cv}} = 1.460$). Mobility patterns are mixed: public transport usage ($M_{\text{PT}} = 18.7\%$) and car ownership remains moderate ($V_{\text{own}} = 526.3$ vehicles/1,000 inhabitants). Despite the balanced urban and economic structure, congestion is substantial ($CL = 32.0\%$), indicating that localized bottlenecks decrease the performance of the road network.

Low-Density Concentrated

4 cities: *Birmingham, Helsinki, Oslo, Toulouse*

The cities in this cluster combine low overall population densities with strongly concentrated economic activities. Population density peaks are the lowest observed across all clusters ($P_{95} = 10,400.2$ inhabitants/km²), while spatial variation remains moderate ($P_{\text{cv}} = 1.196$), reflecting a fragmented and dispersed residential pattern. Road networks are moderately connected ($k_\mu = 4.373$) but suffer from the lowest efficiency across all clusters ($R = 0.377$), suggesting that trips often require longer detours. In contrast to the dispersed population, economic activities are highly centralized: shops ($S_{\text{cv}} = 3.505$) and offices ($O_{\text{cv}} = 2.386$) show the highest spatial variation among multi-city clusters, pointing to strong, economic centers. Despite this centralization, the concentration of traffic flows remains relatively low ($C_{B,95} = 0.016$), indicating that movement remains spread across the network. Mobility patterns are mixed: public transport usage is low ($M_{\text{PT}} = 17.0\%$), car ownership is modest ($V_{\text{own}} = 493.2$ vehicles/1,000 inhabitants), and congestion remains moderate ($CL = 25.8\%$).

Balanced Multimodal

10 cities: *Amsterdam, Brussels, Copenhagen, Frankfurt am Main, Manchester, Paris, Rotterdam, Stockholm, Thessaloniki, Vienna*

The cities in this cluster are characterized by a well-connected road network ($k_\mu = 4.540$) with moderate efficiency ($R = 0.433$), suggesting good accessibility combined with reasonably direct travel paths. Population patterns are balanced, with moderate density peaks ($P_{95} = 17,921.6$ inhabitants/km²) and relatively high spatial variation ($P_{\text{cv}} = 1.268$), indicating more pronounced differences between dense and less dense areas. Economic activities are moderately dispersed, as reflected by the shop ($S_{\text{cv}} = 2.845$) and office ($O_{\text{cv}} = 1.666$) variations. Mobility profiles are strongly multimodal: car ownership is the lowest among all clusters ($V_{\text{own}} = 421.4$ vehicles/1,000 inhabitants), public transport use is moderate ($M_{\text{PT}} = 18.7\%$), and active modes account for a relatively high share of trips ($M_{\text{AM}} = 39.1\%$). Traffic flows are relatively distributed across the network ($C_{B,95} = 0.015$), but congestion levels remain substantial ($CL = 27.1\%$), indicating that even in cities with multiple transport options, there is significant pressure on the road demand.

The seven-cluster configuration captures distinct patterns in network topology, activity distribution, and mobility characteristics across the 32 European cities. Table 5.2 provides a concise overview of the defining characteristics of each cluster.

Table 5.2: Summary of the seven clusters based on road network, activity distribution, and mobility characteristics.

Cluster	Short Description
Centralized Car-Oriented	Strong density contrasts, centralized, car-dominated with low congestion
Homogeneous Car-Oriented	Compact and uniformly dense, car-dominated with low congestion
Dense Multimodal	Extremely dense, multimodal, low car ownership and moderate congestion
PT-Oriented Congested	High public transport use, but also high congestion and car ownership
Well-Connected Bottlenecked	Well-connected but bottleneck-prone, moderate density and mixed mobility
Low-Density Concentrated	Dispersed residential pattern, centralized economic hubs, low PT use
Balanced Multimodal	Mixed densities, strong multimodal mobility, but notable traffic pressure

The spatial distribution of the clusters across Europe, shown in Figure 5.7, reveals clear geographical patterns. Central and Eastern European cities, together with London, form a distinct group characterized by high congestion and strong public transport reliance. Cities with highly centralized yet car-dependent structures, such as Lisbon and Turin, are located predominantly in Southern Europe. The Low-Density Concentrated cities, such as Oslo and Birmingham, are located mainly in Northern and Western Europe, where urban dispersion is more common. Meanwhile, Northwestern European cities such as Amsterdam, Copenhagen, and Paris cluster together into a balanced multimodal group, reflecting a more integrated approach to transport and urban development. Finally, Barcelona remains a unique case within the dataset, combining extreme population density with strong multimodal mobility characteristics, and extreme office centralization. These spatial patterns suggest that geographic, historical, and policy contexts correlate with the cluster results in this research.

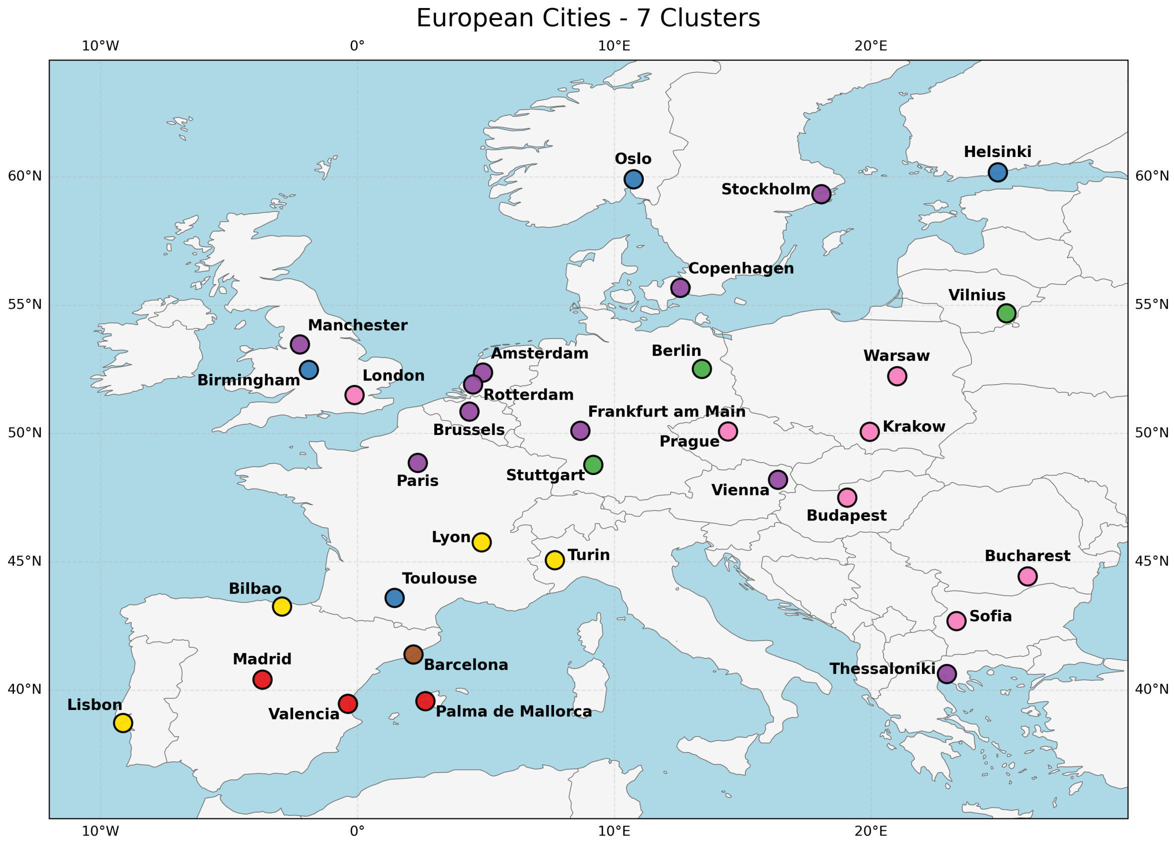


Figure 5.7: Spatial distribution of the seven-cluster result across Europe.

There are some interesting results to be seen in the cluster assignments, especially in terms of geographical and size variation. The three British cities (London, Manchester, and Birmingham) are each assigned to a different cluster, which reflects different urban profiles despite being located in the same country. Thessaloniki is grouped with Northwestern European cities in the Balanced Multimodal cluster, suggesting it shares similar characteristics such as moderate density variation and strong multimodal transport usage, even though it is geographically located in Southern Europe. The Well-Connected Bottlenecked cluster includes a very large city and two smaller cities, bringing together Berlin, Stuttgart, and Vilnius. This indicates that differences in scale do not necessarily prevent cities from sharing comparable network structures and traffic flow characteristics. It also highlights the importance of considering local context before transferring solutions between cities in the same cluster.

5.2.2. Two-Cluster Result

The two-cluster result, for which relatively high silhouette scores were computed across all three methods, demonstrates a high degree of consistency between K-Means, K-Medoids, and Ward's method. The ARI values, presented in Figure 5.8, confirm strong agreement across all method combinations: K-Means and K-Medoids achieve an ARI of 0.86, K-Means and Ward's method 0.87, and K-Medoids and Ward's method 0.74. These results indicate that, despite methodological differences, the dataset supports a robust and stable two-cluster classification. This outcome is somewhat unexpected, given the variation in indicator values previously observed for the four example cities discussed in Section 4.1.

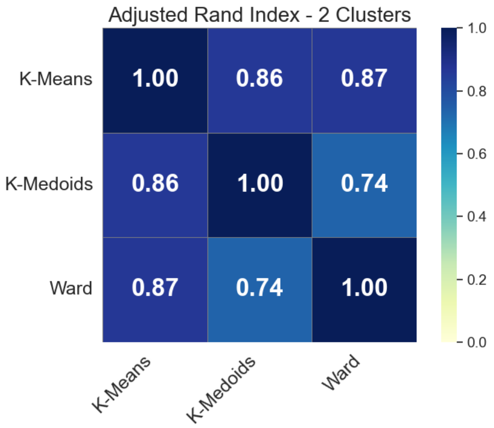


Figure 5.8: Adjusted Rand Index comparing the two-cluster results obtained from K-Means, K-Medoids, and Ward's method.

The geographical distribution of the two clusters, illustrated in Figure 5.9, reveals a clear regional separation. Cities from Spain and Italy consistently group into one cluster, while the remaining cities, predominantly located in Northern, Central, and Eastern Europe, form the second cluster. Lisbon and Lyon, which are not consistently assigned across methods, are indicated separately in gray. This division suggests that structural and mobility characteristics in Southern Europe differ significantly from those observed in the rest of Europe, a pattern that is also visible in Figure 5.4, where the southern cities merged with the other cities at the highest linkage distance.

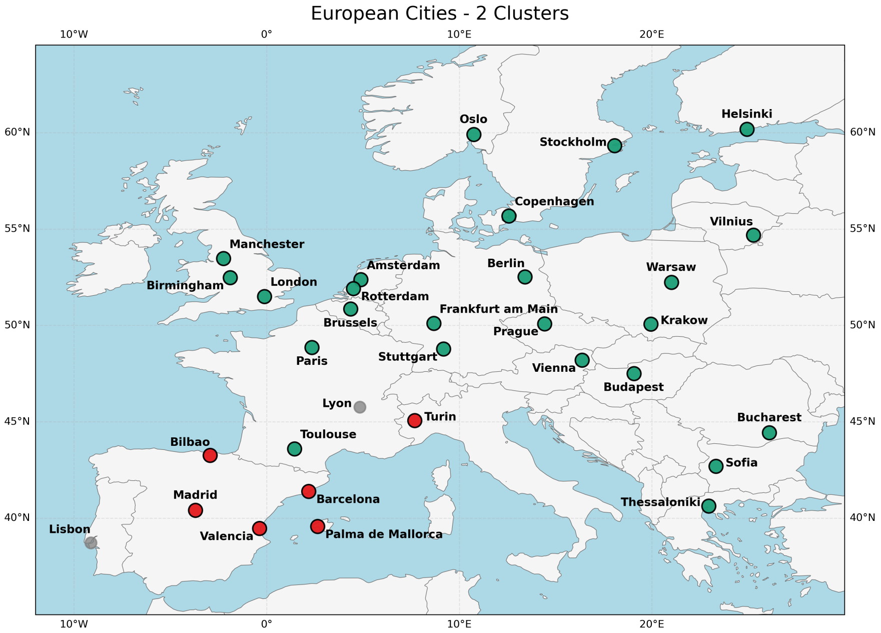


Figure 5.9: The 32 European cities grouped into two clusters. Lisbon and Lyon are not clustered, but illustrated in gray.

When compared to the seven-cluster result, the two-cluster configuration provides an interesting perspective. The southern cities, joined by Lisbon and Lyon, correspond to the clusters labeled as Centralized Car-Oriented, Homogeneous Car-Oriented, and Dense Multimodal. In contrast, the remaining cities, grouped into PT-Oriented Congested, Well-Connected Bottlenecked, Low-Density Concentrated, and Balanced Multimodal, form the second cluster. This division highlights strong structural differences: the southern cities exhibit higher peak population densities, lower congestion levels, and a stronger reliance on car travel, whereas the second group tends to have higher road network connectivity, lower car ownership, and greater public transport usage. A full overview of the characteristics of the two clusters can be found in Section D.1.

5.2.3. Five-Cluster Result

The five-cluster configuration provides a more detailed classification of the 32 cities, but shows greater instability across the clustering methods compared to the two-cluster result. As visualized in the ARI heatmap in Figure 5.10, K-Means and Ward's Method show relatively high agreement (ARI = 0.80), indicating that these two methods identify a similar structure when grouping the cities into five clusters. In contrast, K-Medoids shows considerably lower agreement with both K-Means and Ward's Method (ARI = 0.43 in both cases), indicating a substantially different clustering result. These results confirm that the five-cluster result is less robust across methods, primarily due to variation introduced by K-Medoids. This difference is also anticipated by the silhouette scores discussed in Section 5.1, where K-Medoids showed weaker cluster quality at $k = 5$.

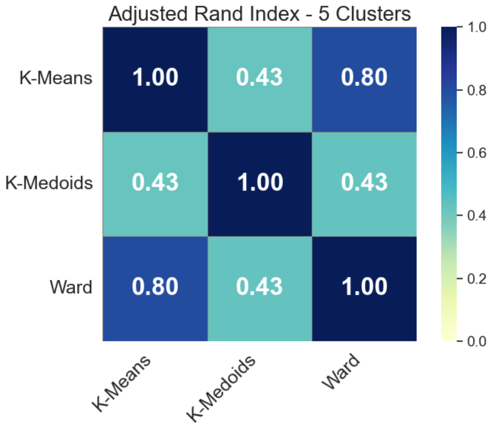


Figure 5.10: Adjusted Rand Index comparing five-cluster results for K-Means, K-Medoids, and Ward's Method.

The spatial distribution of the five-cluster result is shown in Figure 5.11. As the ARI values suggest, the clustering results do not align across the three methods, resulting in a larger number of cities that could not be consistently assigned to a single cluster. Lisbon and Lyon remain inconsistently assigned, while Palma de Mallorca and Vienna also switch cluster assignments depending on the method. These cities appear to occupy positions between certain clusters, displaying characteristics that overlap with these groupings.

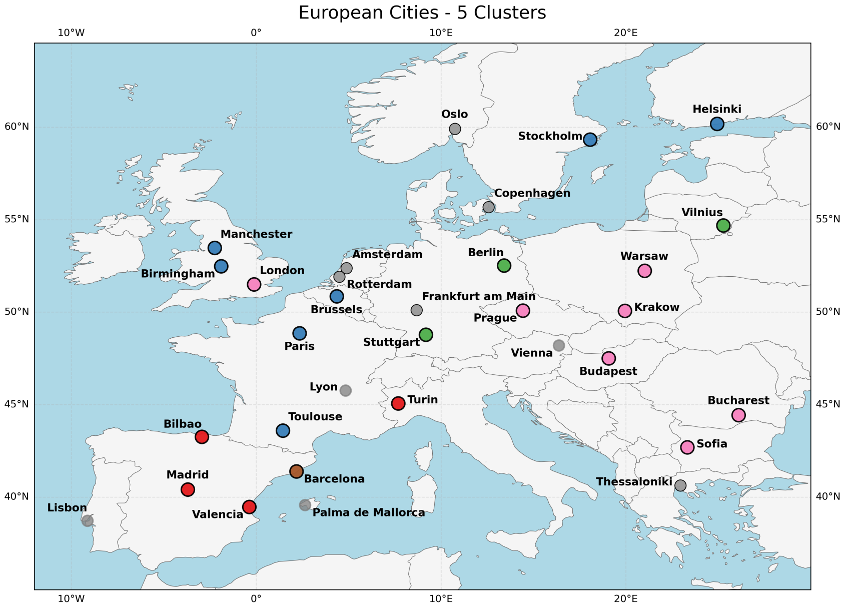


Figure 5.11: The 32 cities grouped into five clusters. Inconsistent assigned cities are shown in gray (Lisbon, Lyon, Palma de Mallorca, Vienna) or also outlined as group (Amsterdam, Copenhagen, Frankfurt am Main, Oslo, Rotterdam, Thessaloniki).

Notably, a group of cities (Amsterdam, Copenhagen, Frankfurt am Main, Oslo, Rotterdam, and Thessaloniki) consistently appear together across all three methods but is assigned to two different clusters: either the cluster containing Berlin, Stuttgart, and Vilnius or the cluster containing the other Northwestern cities, depending on the algorithm. This suggests that while these cities are internally similar, the methods differ in how they position this group within the broader cluster structure. Their inconsistent assignment is the cause of the observed variation in ARI scores. More importantly, this inconsistency indicates that a five-cluster configuration does not provide a fully distinct or stable division of European cities.

A possible explanation for the instability is the sensitivity of K-Medoids to the selection of medoids, particularly in small datasets. In K-Medoids, cluster centers (medoids) are located at actual data points. Although K-Medoids is theoretically more robust to outliers than K-Means, the small sample size in this study amplifies its sensitivity to the exact location of individual data points. This sensitivity likely contributes to the observed instability at $k = 5$. By contrast, K-Means and Ward's Method, which are not restricted to existing data points for their cluster centers, produce more similar groupings. Nevertheless, the differences between the methods help uncover additional patterns within the dataset. Even when cluster assignments shift, these variations reveal underlying structural similarities and differences among cities, contributing valuable insights to the overall analysis.

The five-cluster result already defines three of the clusters that appear in the more detailed seven-cluster result. The Dense Multimodal cluster, consisting solely of Barcelona, emerges consistently as a distinct group, reflecting its combination of extremely high population density and significant variation in office density distribution. Similarly, the PT-Oriented Congested cluster, composed of Central and Eastern European cities together with London, is already clearly formed, capturing the pattern of high congestion levels and car ownership, and strong reliance on public transport. The Well-Connected Bottlenecked cluster, including Berlin, Stuttgart, and Vilnius, is also defined in the five-cluster configuration. These stable groupings show the robustness of the underlying urban contexts of these clusters.

In contrast, the remaining clusters are not clearly defined at $k = 5$. The cluster containing the southern cities, which includes Bilbao, Madrid, Turin, and Valencia, later divides (together with inconsistently assigned cities) into the Centralized Car-Oriented and Homogeneous Car-Oriented clusters at $k = 7$, indicating internal heterogeneity that only becomes apparent with more detailed classification. Similarly, the cluster with northwestern cities, composed of cities such as Brussels, Paris, and Stockholm, remains broader in the five-cluster result and subsequently splits, together with the Inconsistently Assigned group, into two distinct clusters at $k = 7$. This difficulty in cleanly separating the Northwestern cities at $k = 5$ is consistent with the instability observed across clustering methods. Overall, the five-cluster result provides a meaningful initial structuring of the European cities, while simultaneously highlighting the degree of variability in certain urban patterns at this clustering level. The full overview of the characteristics of the five clusters and the Inconsistently Assigned group can be seen in Section D.2.

5.2.4. Performance of Clustering Methods

In addition to the Adjusted Rand Index comparison between the three clustering methods, the silhouette scores for the two-, five-, and seven-cluster configurations are compared to assess the performance of the methods. These results are summarized in Table 5.3. For the seven-cluster configuration, all three methods show identical silhouette scores, as expected given that they produce identical cluster assignments. In the two-cluster configuration, both K-Means and K-Medoids achieve higher silhouette scores than Ward's Method. This difference, consistent with the earlier analysis in subsection 5.2.2, reflects the assignment of Lisbon and Lyon, which are grouped with Northwestern and Eastern cities in K-Means and K-Medoids, rather than with Southern cities as in Ward's Method. For the five-cluster configuration, K-Medoids performs significantly worse, with a noticeably lower silhouette score, as already highlighted in subsection 5.1.1. In contrast, K-Means achieves the highest score at this level, while Ward's Method shows a silhouette value similar to that of the seven-cluster result.

Taken together, these results suggest that K-Means offers the most consistently high silhouette scores across the tested configurations, indicating a relatively stable ability to form internally cohesive and well-separated clusters. Ward's Method also performs robustly, especially in the five- and seven-cluster

Table 5.3: Silhouette scores for each clustering method at 2, 5, and 7 clusters.

	K-Means	K-Medoids	Ward's Method
2 clusters	0.281	0.293	0.252
5 clusters	0.283	0.201	0.267
7 clusters	0.267	0.267	0.267

solutions, and shows strong alignment with K-Means based on ARI values. K-Medoids, on the other hand, appears more sensitive to the number of clusters: although it performs well for two clusters, its silhouette score drops substantially at five clusters, suggesting less coherent groupings and greater sensitivity to medoid selection.

This chapter has examined the clustering results based on the included topology, economic activity, and mobility characteristics of the 32 cities. Across K-Means, K-Medoids, and Ward's method, the configurations with two, five, and seven clusters were evaluated in detail. Among these, the seven-cluster configuration proved to be the most robust and informative, offering a distinct characterization of urban contexts. Meanwhile, the two- and five-cluster configurations offer valuable complementary perspectives, revealing geographical trends and interesting groupings. Together, these findings provide a solid foundation for the discussion in Chapter 6 and the conclusions drawn in Chapter 7.

6 Discussion

This chapter discusses the findings of the city characterization and clustering analysis. First, Section 6.1 summarizes the main results from Chapter 4 and Chapter 5, offering an initial reflection on the observed urban patterns and clustering results. These outcomes are then placed within the broader context of urban and transport research in Section 6.2, comparing them to insights from previous studies. Finally, Section 6.4 critically goes into the methodological limitations of this research and their implications for the robustness and interpretation of the results. Throughout the discussion, the focus is on interpreting how the findings contribute to the understanding of urban transport dynamics and the opportunities they offer for cross-city learning.

6.1. Main Findings

The characterization of the 32 cities revealed different forms of road network structure, population and economic activity distribution, and mobility patterns. Indicators related to network topology, such as the average node degree and network efficiency, showed relatively limited variation across the European cities, possibly reflecting general structural features of urban road networks. An exception was the 95th percentile of betweenness centrality, which exhibited substantial variation and highlighted differences in the concentration of traffic flows across networks. In contrast, indicators describing population density, the distribution of shops and offices, and mobility patterns showed more variation. Measures capturing the intensity and spatial dispersion of population and facilities emphasized the contrast between highly centralized and more evenly distributed urban forms. Mobility indicators similarly revealed a wide range of profiles, from car-dependent to multimodal cities, and from low to high levels of congestion. Among these, the share of trips made by public transport showed particularly high variability. Overall, the selected indicators offer a comprehensive description of urban structure and network performance, forming a strong foundation for distinguishing cities through clustering analysis.

To explore relationships between the seventeen initial indicators, a correlation analysis was performed. Most indicators showed only weak correlations, suggesting that they captured different aspects of urban structure and mobility. However, some unexpected patterns emerged. Indicators related to node degree were negatively correlated with both the mean and the 95th percentile of population density. This was not directly anticipated based on the literature and may indicate that highly connected road networks are not necessarily associated with dense urban forms. Another notable finding was the absence of significant correlations between congestion levels, car ownership, and the share of trips made by motorized vehicles. This suggests that higher car ownership does not automatically lead to greater car use or traffic congestion, pointing instead to other factors such as wealth levels or traffic management effectiveness. These findings challenge conventional transport planning assumptions, where car ownership is often seen as a measure for car dependence or congestion. The weak relationships observed here highlight the need to analyze structural and behavioral indicators in combination rather than alone. Urban form, local policies, and income levels may all seem to influence whether people actually use their cars, showing that car ownership alone does not explain mobility patterns in European cities.

Where strong overlaps were observed, such as between the mean and 95th percentile of population density, or between the mean and variability of shop and office densities, indicators that better captured spatial distribution were prioritized. This selection narrowed the set to ten indicators, and after applying Principal Component Analysis, 79.3% of the total variance was captured by five principal components. Building on this reduced data structure, the three clustering methods (K-Means, K-Medoids, and Ward's Method) offered clear insights into the clustering structure of the cities. Across all methods, configurations with two, five, and seven clusters emerged as the most meaningful, based on silhouette analysis, elbow plots, and dendrogram interpretation.

The two-cluster result revealed a strong geographical pattern. Southern European cities, particularly those from Spain and Italy, contrasted sharply with cities in other parts of Europe in terms of network structure, activity distribution, and mobility behavior. Although the assignment of Lisbon and Lyon was less clear, the division remained highly consistent across methods, indicating that regional context influences urban form and transport systems at this level of analysis.

The five-cluster result provided a different classification, attempting to subdivide the two larger groups identified at the two-cluster level. However, it struggled to do so consistently across methods. City assignments, particularly between K-Medoids and the other approaches, were often unstable. While certain groups, such as the Central and Eastern European cities characterized by strong public transport reliance, were consistently identified as a cluster, other groupings showed shifting assignments across methods. This instability suggests that five clusters could not fully subdivide the variability present in the dataset for the 32 cities.

In contrast, the seven-cluster result proved to be both the most stable across methods and the most informative outcome. Full agreement indicated a robust and well-structured configuration. The seven clusters captured several patterns in urban contexts, highlighting meaningful differences in network structure, activity distribution, and mobility characteristics. Some clusters grouped cities characterized by high congestion levels combined with strong public transport reliance, while others reflected highly centralized, car-oriented cities with dense cores but lower congestion. Additional clusters distinguished cities with low-density residential patterns and concentrated economic hubs from those where both population and activities were more evenly distributed. Differences in car ownership levels and the degree of reliance on non-car modes further emphasized the distinctiveness of the groups. As the most notable result, Barcelona stood out as a singular case, reflecting its extreme population density and high concentration of office activity, setting it apart from all other cities in the dataset.

Reflecting on these findings, several broader patterns emerge. First, the absence of strong correlations between car ownership, motorized travel, and congestion levels suggests that these dimensions are shaped by different underlying factors. High car ownership does not necessarily translate into greater car use or congestion, as it may reflect factors such as wealth or lifestyle. Congestion, meanwhile, often depends more on network structure and traffic management. This is evident in the two-cluster result, where Southern European cities combine high car ownership with relatively low congestion.

Second, the clustering results reveal a consistent regional pattern: cities in Southern Europe differ clearly from their Northern and Eastern European counterparts in terms of network structure, activity distribution, and mobility characteristics. This separation likely reflects deeper historical and institutional differences. Whereas Northern and Central European cities often developed through structured urban planning or post-war reconstruction, Southern European cities often retained compact historical centers and more organically developed urban forms. These layouts result in higher central densities and more localized activity patterns. In addition, differences in transport policy, such as investment levels and traffic management strategies, are likely contributing to contrasting mobility outcomes. Climate conditions may also influence walkability and cycling, either encouraging active modes or discouraging them during heat periods. Together, these long-standing development trajectories help explain why Southern cities consistently form a separate cluster across methods.

Third, despite the complexity of urban transportation systems, meaningful groupings can still be identified when cities are compared across multiple domains. The stability of the seven-cluster result demonstrates that cities with shared structural and behavioral characteristics can be grouped meaningfully, even when drawing on a broad set of indicators. This reinforces the value of multidimensional classification frameworks and offers promising opportunities for comparative research and cross-city learning.

6.2. Comparison to Previous Research

Earlier studies have shown that clustering cities based on transport network characteristics can reveal structural differences in connectivity and centrality. For example, Tundulyasaree (2019) and Yamaoka et al. (2021) used graph-theoretical indicators to group cities according to network topology, while Badhruudeen et al. (2022) focused on geometric features such as grid-like versus organic layouts. These studies demonstrated clear geographic patterns in urban form and emphasized the value of network-based classification.

The findings of this study support those conclusions but also offer several refinements. While indicators such as average node degree and the network efficiency ratio were relatively uniform across the 32 European cities, the 95th percentile of betweenness centrality ($C_{B,95}$) showed substantial variation. This confirms earlier observations that $C_{B,95}$ captures flow concentration and structural hierarchy in urban networks (Louf & Barthelemy, 2014; Strano et al., 2013). At the same time, the limited variation in other topological indicators suggests the presence of underlying regularities in network form, as found by Badhruudeen et al. (2022), yet highlights the risk of overlooking functional differences when relying on a smaller indicator set. By incorporating multiple dimensions of network structure, this research offers a broader foundation for clustering.

Beyond physical structure, previous research has emphasized land-use composition and mobility behavior as key dimensions of urban classification. Studies such as Puissant and Eick (2024) and Coenegrachts et al. (2024) have clustered cities based on facility distribution and shared mobility markets, respectively. These contributions underscore that spatial organization and travel behavior shape urban dynamics in distinct but complementary ways.

This research confirms that perspective but distinguishes itself by integrating network topology, facility dispersion, and modal preferences into a single clustering framework. Whereas earlier studies typically focused on one domain, the combined approach adopted in this research reveals underlying patterns. For instance, cities with similar population densities but different levels of car ownership and public transport usage. Although the domains of activity distribution and mobility patterns do not show strong correlations, their interaction becomes meaningful when considered in combination with network topology. The emergence of stable clusters illustrates that the integration of structural, functional, and behavioral indicators offers a more holistic explanation than domain-specific classifications alone.

While some earlier studies, such as Tundulyasaree (2019), applied more than one clustering method for validation, most rely on a single technique. In this study, K-Means, K-Medoids, and Ward's hierarchical clustering were applied in parallel to evaluate the robustness of the cluster results. The full agreement across all three methods for the seven-cluster result strengthens confidence in the identified typologies. This level of cross-method stability suggests that the observed groupings reflect meaningful structural and behavioral differences rather than methodological characteristics.

In contrast to studies that isolate either urban form or transport behavior, this research emphasizes their interplay. For example, while high car ownership is often associated with higher congestion in mobility-focused classifications, the results here show that this relationship is not consistent across the 32 included cities. Similarly, the recurring grouping of Southern European cities across clustering methods reflects shared structural and behavioral characteristics and supports earlier findings, but now placed within a broader, integrated framework.

In summary, this research confirms and extends earlier findings on transport networks, land use, and travel behavior. By integrating road topology, population density, economic activity dispersion, and modal preferences into a single clustering framework, it enables a more holistic characterization of European cities. This multidimensional approach strengthens comparative urban analysis and supports more nuanced insights into how structure and behavior shape urban systems.

6.3. Transferability & Application

The clustering results offer a practical foundation for cross-city learning by identifying groups of cities with comparable structural and mobility conditions. Interventions such as investments in public transport, active mode infrastructure, or road network optimization often involve complex trade-offs and require contextual justification. When cities face similar challenges, learning from their counterparts can reduce uncertainty, accelerate implementation, and improve policy effectiveness. In this sense, clustering enhances the evidence base for urban policy by identifying structurally relevant reference cases, supported by data and strengthening informed and adaptive strategy development.

One example is the PT-Oriented Congested cluster, which includes Central and Eastern European cities such as Budapest, Warsaw, and Prague, as well as London. Despite differing national and economic contexts, these cities all face the shared challenge of managing high congestion levels alongside strong public transport usage and relatively high car ownership. What unites them further is their role as national capitals and regional hubs. As major employment, education, and administrative centers, they attract large numbers of commuters from surrounding areas, often far beyond their municipal boundaries. This intensifies peak-hour pressure on both road networks and public transport systems. In such contexts, strategy exchange may focus not only on internal network optimization but also on managing metropolitan-scale demand, such as park-and-ride systems, regional rail integration, or congestion pricing. The effectiveness of such strategies can then be used for potential integration in other cities within the cluster.

Cross-cluster comparison can also highlight contrasting profiles and complementary lessons. Amsterdam, part of the Balanced Multimodal cluster, features low car ownership, moderate congestion, and strong use of active modes. In contrast, cities like Turin or Lisbon in the Centralized Car-Oriented cluster exhibit higher car dependency and less balanced activity patterns. If a city like Turin seeks to promote cycling or reduce car dominance in dense cores, Amsterdam offers an example of how similar urban densities can be paired with strong active mode strategies, such as dedicated cycle networks, traffic calming zones, and pedestrianized areas. Although not in the same cluster, such examples can support context-aware adaptation strategies aligned with specific policy objectives.

The classification also provides guidance for cities beyond the original dataset. For instance, a city like Belgrade, which features a dense historical center, rising car ownership, and a relatively well-developed public transport network, may share key characteristics with the PT-Oriented Congested cluster. Although not part of the 32 included cities, Belgrade faces many of the same urban challenges as cities like Budapest or Warsaw. By comparing indicator profiles, policymakers can identify relevant cities, strategies, and collaborate on solutions to manage congestion, improve public transport, or coordinate regional travel demand.

Beyond guiding immediate strategy development, the clustering framework also offers a foundation for long-term, adaptive urban policymaking. As cities continue to evolve, the classifications can support benchmarking, allowing policymakers to track progress, evaluate the effects of interventions, and anticipate emerging challenges. Repeating the clustering analysis with updated data enables cities to reflect on whether they are shifting toward more balanced, efficient, or sustainable profiles, or diverging from their intended trajectories. The results serve not only as a snapshot of current urban transportation characteristics, but as a tool to inform ongoing decision-making, align policy with structural change, and support more resilient, future-oriented planning across diverse urban contexts.

6.4. Limitations & Implications

While the analysis offers valuable insights into the urban form of European cities, several limitations affect how the findings should be interpreted. These limitations arise from methodological choices, data availability, and practical constraints. This section provides an overview the most relevant limitations and discusses their implications for the robustness and comparability of the results.

Number of Cities

An important limitation of this research lies in the relatively small number of cities included. While the 32 cities selected represent some of the most significant urban areas in Europe, they account for only a fraction of the 828 cities across the continent, let alone globally. As discussed in Section 3.1, the limited sample was primarily constrained by data availability, particularly the need for consistent and comparable indicators across the considered domains.

This restricted sample size has several implications. With more cities, additional patterns might emerge, and certain statistical relationships could become more robust. Correlations between indicators could be confirmed with greater confidence, while the risk of coincidental findings would decrease. A larger set could also lead to the identification of new clusters or offer clearer groupings for cities that currently appear as outliers, such as Barcelona, which may share characteristics with cities not included in the present dataset. However, expanding the dataset would also introduce challenges. Including more cities would make the analysis more resource-intensive, potentially affecting feasibility in terms of the collection, processing, and interpretation of data. It may also reduce the level of detail with which individual cities can be considered. Balancing depth and wide application is therefore an important consideration when applying this approach to wider sets of cities.

City Boundaries

Another limitation relates to the way in which urban boundaries were defined. As discussed in Section 2.1, the delineation of urban areas is fundamentally subjective, depending on the research objective and context. The approach adopted in this research, which uses the area within the main ring road as a basis, introduces uncertainty, particularly for cities with multiple ring roads (e.g. Paris, Madrid, London) or those without a clear ring structure (e.g. Vilnius, Stuttgart). This raises questions about which suburban areas were or should have been included or excluded from the analysis.

These boundary decisions have implications for the comparability of results across cities. Small differences in how the boundary is drawn can influence the number of nodes, the extent of the network, and the spatial distribution of facilities and population. Additionally, the boundaries used for certain indicators (mobility and congestion level) were not clearly defined in the original data sources. As a result, there is likely a small mismatch between the geographical area for the road topology, population, and economic activity indicators on one hand, and the mobility indicators on the other. This reduces the consistency within the dataset.

To enhance reproducibility, the initial boundary polygons are included in the supplementary materials attached alongside this report. The additional constraint, iteratively including nodes within 200 meters of the polygon boundary, helped improve the delineation of the study area and reduce the subjectivity of the initial polygons. Nevertheless, more precise and harmonized boundary definitions would likely improve the quality and reliability of the results. This was considered too time-consuming for the scope of this project.

Data from Different Years

A further limitation of this research is that the data across domains was collected in different years. As noted in Table 3.2, the dataset spans a period from 2020 to 2025, reflecting the most recent available data for each domain. While this ensures up-to-date input, it also introduces a degree of misalignment between domains. Collecting all indicators from the same reference year would provide a more consistent basis for analysis and likely improve the accuracy of cross-city comparisons.

This variation is not expected to significantly affect the findings, as the research focuses on structural characteristics such as road networks, population distribution, and general mobility behavior, which typically change slowly over time. However, it may offer a slightly less precise view of the situation in certain cities. This is particularly relevant for mobility indicators: modal shares for four cities (Bilbao, Budapest, Thessaloniki, and Vienna) were drawn from an older report, and car ownership figures for three other cities (Frankfurt am Main, Lisbon, and Sofia) were obtained from alternative publications due to missing data in the main sources.

OpenStreetMap Data Quality

The limitations of OpenStreetMap data, previously noted in subsection 3.2.3, are further elaborated here due to their impact on facility indicators. Since OpenStreetMap is a publicly sourced platform maintained by a large community of contributors, consistency in data quality across cities cannot be guaranteed. While this is unlikely to significantly affect the road network data, as streets are relatively well-defined and uniformly mapped, it does create challenges for uniformly classifying facilities. The tagging of shops, offices, and other urban activities can vary between countries, cities, and contributors and often overlaps. For instance, the distinction between shops and amenities, or between offices and industrial uses, is not always clearly defined. Moreover, many amenities and shops also function as workplaces, further complicating the categorization.

To address this uncertainty, average values based on absolute facility counts were given lower priority in this analysis. Instead, the analysis focused on distribution patterns using the coefficient of variation, which is more robust to inconsistencies in documentation, as it captures relative differences within each city rather than relying on absolute counts between cities. This approach assumes that classification practices are relatively consistent within each individual urban area, allowing meaningful spatial patterns to be observed despite cross-city differences in documentation practices. Nevertheless, more detailed and standardized facility data would further improve the accuracy and comparability of these indicators.

Interrelations in Activity Distribution

While distribution indicators for population, shops, and offices were included separately, their combined spatial configuration, often linked to urban polycentricity, was not explicitly analyzed in this research. Although initial patterns can be observed by comparing their separate distributions across the urban area, no systematic method was applied to assess the spatial overlap or alignment between these activity types. As a result, potentially important relationships between mixed-use areas, the clustering of activity centers, and mobility behavior may not have been fully captured.

7 Conclusion

This chapter presents the main conclusions of the research. Section 7.1 first answers the main research question, followed by the sub-questions in Section 7.2, based on the results of the city characterization and clustering analysis. Section 7.3 highlights the scientific contributions of this research. Recommendations based on the findings are provided in Section 7.4, and opportunities for future research are provided in Section 7.5.

7.1. Main Research Question

The objective of this research was to explore how urban characteristics related to road networks, activity distribution, and mobility patterns can be used to cluster European cities into meaningful groups. To achieve this, the research applied multiple clustering methods to a selected set of indicators across these domains, answering the main research question:

How can the application of multiple clustering methods reveal distinct groups of European cities based on road network, activity distribution, and mobility characteristics?

The findings show that for the 32 included European cities, a clear and meaningful classification into seven distinct clusters could be achieved. These clusters captured road topology, activity distribution, and mobility characteristics, which determined distinct city profiles.
The use of three different clustering methods, K-Means, K-Medoids, and Ward's Method, demonstrated that clustering outcomes were relatively consistent across the approaches for two and seven clusters. In particular, the seven-cluster result showed full agreement across all methods, highlighting the robustness of the clustering result. Interpretable patterns also emerged for the two- and five-cluster results, although with less stability for the five-cluster result. Applying multiple methods strengthened the confidence that the identified groups represent meaningful differences between cities rather than being influenced by the choice of clustering method.

Thus, this research concludes that a clustering approach considering multiple domains provides a strong foundation for distinguishing cities based on their combined structural and mobility characteristics. It offers valuable opportunities for comparative research and cross-city learning, helping cities to better understand shared challenges and potential pathways for development.

7.2. Sub-Questions

To further structure the conclusions, this section answers the four sub-questions formulated for this research. Each sub-question addresses a specific aspect of the methodology or findings, contributing to a comprehensive understanding of how cities can be distinguished based on their structural and mobility characteristics.

Indicators to Distinguish Cities

Which indicators most effectively distinguish cities in terms of road topology, activity distribution, and mobility characteristics?
How can these indicators be quantified and compared across European cities?

To effectively distinguish cities, indicators were selected that capture structural, spatial, and behavioral characteristics. For road network topology, the average node degree, the efficiency ratio, and the 95th percentile of betweenness centrality quantified differences in connectivity, efficiency, and the robustness of the network. For activity distribution, the coefficients of variation for population density, shops, and offices captured how unevenly these functions were distributed across urban areas, while the 95th percentile of population density described the density in the densest neighborhoods. Mobility characteristics were measured through modal shares (motorized vehicles, public transport, active modes), car ownership rates, and congestion levels.
Indicators were either calculated based on publicly available data sources or collected directly from public reports and databases. To enable consistent comparisons between cities, individual indicators were standardized as needed: betweenness centrality indicators were normalized to account for variation in city size, and coefficients of variation were used to capture relative distributions of facilities and population within each city. Following this, all indicators were Z-score normalized to allow meaningful comparisons across different indicators with varying scales. Among all indicators, the variation in betweenness centrality, the dispersion of economic facilities, and the share of trips made by public transport proved most effective in distinguishing urban contexts.

Relationships Between Indicators

What relationships exist between the included indicators, and what do these relationships reveal about underlying urban dynamics?

The analysis revealed that most indicators captured distinct aspects of urban structure and mobility, as only a few strong cross-domain correlations were observed. Where strong relationships existed, they mainly reflected expected overlaps, such as between the mean and 95th percentile of population density, and between mean and variation-based indicators for shops and offices. These findings justified the prioritization of distribution indicators, such as percentiles and coefficients of variation, to better differentiate between cities.
Beyond overlapping patterns, the observed correlations also highlighted urban dynamics. For example, cities with higher shares of active mobility generally experienced lower congestion levels. These relationships indicate that structural, functional, and behavioral characteristics are interrelated, reinforcing the importance of considering multiple domains together when analyzing and comparing cities.

Clustering Methods and Their Application

Which clustering methods are most suitable for grouping European cities based on these indicators, and what are their respective advantages and limitations?

The research showed that applying multiple clustering methods strengthens the robustness of city groupings. K-Means, K-Medoids, and Ward's Method all proved effective in identifying meaningful patterns across a small set of cities, despite their methodological differences. K-Means was computationally efficient and produced stable results across hundreds of runs. K-Medoids offered greater robustness to outliers but showed some sensitivity to the location of individual data points within the dataset. Ward's Method provided valuable insights into how clusters merge hierarchically, without requiring a predefined number of clusters. However, relying solely on Ward's Method would make interpretation more challenging, as dendrograms become increasingly difficult to analyze in detail when the number of cities grows.

Consistency of Clustering Outcomes

How consistent are the clustering outcomes across different methods, and how can their similarities and differences be interpreted?

The clustering outcomes showed a high degree of consistency across K-Means, K-Medoids, and Ward's Method, for the two- and seven-cluster results. The strong agreement at the two-cluster

level reflected a clear regional division between Southern cities and the rest of Europe, while the full agreement in the seven-cluster result confirmed that the 32 European cities could be grouped in distinct and stable urban typologies, independent of the clustering technique used.
For the five-cluster result, the cluster configurations differed significantly between K-Medoids and the other two methods. These inconsistencies, however, offered valuable insights into cities that lie between groups or show more mixed characteristics. Using multiple clustering methods thus not only strengthened the robustness of the identified groupings but also helped uncover alternative perspectives on the diversity among cities, enhancing the overall reliability of the results for comparative urban analysis.
Among the included methods, K-Means and Ward's Method emerged as the most consistent given their silhouette scores and high agreement across cluster configurations. While K-Medoids showed greater variability, particularly at five clusters, it contributed complementary insights that enriched the interpretation of cluster structures.

7.3. Scientific Contributions

This research contributes to the field of urban transport and spatial analysis in several important ways, both by addressing a research gap and by demonstrating a systematic methodological approach.

First, it addresses a gap in the existing literature by integrating road network structure, activity distribution, and mobility characteristics into a single clustering framework. Whereas previous studies have classified cities based on individual domains such as road topology, land use, or shared mobility, this research combines the structural, functional, and mobility characteristics of urban areas. This approach provides a more holistic classification of European cities and highlights the interdependencies between urban form, activity patterns, and mobility systems.

Second, this research demonstrates that applying multiple clustering methods strengthens the robustness and interpretability of city classifications. By systematically comparing K-Means, K-Medoids, and Ward's Method, it shows that meaningful and stable urban typologies can be identified independently of the chosen algorithm. The strong agreement across methods, particularly for the seven-cluster result, indicates that the groupings reflect real differences between cities, even when considering a wide range of structural, functional, and behavioral characteristics.

Third, this research offers a comprehensive and systematic methodology for classifying cities by combining public data sources, multiple clustering algorithms, and complementary evaluation metrics. This integrated approach strengthens the robustness, comparability, and replicability of the findings, providing a solid foundation for future cross-city analyses and accelerating collective solutions to urban challenges, even when working with smaller datasets.

Ultimately, clustering cities based on topology, activity distribution, and mobility can provide a solid foundation for addressing shared challenges, not by reinventing the wheel, but through cooperation, cross-city learning, and a more collaborative, data-based approach to shaping the future of urban transport systems.

7.4. Recommendations

This section proposes ways to strengthen the methodology and expand the future application of the clustering framework. The recommendations focus on extending the dataset to capture different stages of urban development and on improving the quality and coverage of input data, while maintaining the systematic and replicable approach introduced in this study.

Development Timestamps

This research focused on comparing European cities within the same time frame. A valuable extension would be to include multiple profiles for each city that reflect different stages of their development. For example, rather than a single entry for Amsterdam, the dataset could include "Amsterdam (2005)", "Amsterdam (2015)", and "Amsterdam (2025)", capturing how the city has evolved in terms of road network structure, activity distribution, and mobility behavior.

Such an approach would enable cities not only to find comparable cities in the present, but also to explore development trajectories over time. A city in 2025 that is similar to “Amsterdam (2005)” could study the strategies introduced during that period and evaluate whether they effectively addressed the challenges by 2015, or whether they produced long-term effects, positive or negative, by 2025. In this way, the clustering method becomes a more dynamic tool for learning from past transformations and identifying strategies suitable for specific stages of urban transportation development. It can also help individual cities focus more on their own development history.

From a research perspective, the inclusion of historical city profiles would make it possible to study how urban form and transportation systems evolve together. It could also uncover shifting relationships between certain indicators as cities develop, offering more nuanced insights into the drivers of urban change and supporting long-term policymaking.
Given the need for large volumes of data, building such longitudinal profiles would be resource-intensive. It requires access to consistent historical data, standardized indicator definitions across countries, and careful processing. However, the potential benefits for both researchers and decision-makers are considerable. In the context of the European Union, streamlining such data collection may be a feasible long-term objective. The methodology presented in this study could be directly applied to longitudinal data, offering a practical framework for understanding urban and transport development over time.

Improving Dataset Quality and Coverage

There are several ways to improve the quality of the dataset, which strengthens the reliability of future clustering results. Increasing the precision of urban boundary definitions, improving the consistency of input data across indicators, and ensuring closer alignment between the indicators and the actual characteristics of each city would allow the unique context of each individual city to be captured more accurately. Collaboration with local authorities and municipalities would assist this process. More precisely delineated boundaries, consistently applied across all measured indicators, would reduce inconsistencies in spatial analysis and enhance comparability between cities. Similarly, using more complete and systematically validated data sources, particularly for facility locations (POI-data) and mobility characteristics, would minimize data inaccuracies and better represent the functional reality of each city. Together, these improvements would enable the clustering results to reflect the specific local context that influences urban development and mobility patterns. Importantly, all of these enhancements can be integrated within the same methodological framework presented in this study, preserving comparability while improving the robustness and interpretability of the clustering results.

If such improvements in data quality could be achieved across a larger sample of cities, the explanatory strength of the analysis would significantly increase. Especially if this methodological approach is applied to different continents. A more extensive dataset would increase the robustness of observed relationships between indicators, reduce the influence of outliers, and allow the identification of more nuanced urban patterns. Moreover, it would strengthen the potential for cross-city learning by offering a richer set of comparable cases, enabling cities to learn more precise lessons from others with similar development stages. In doing so, the clustering framework would become an even more powerful foundation for understanding urban dynamics and supporting tailored policy interventions.

7.5. Future Research

Building on the findings, contributions, and limitations of this research, several directions for future research are proposed. These suggestions aim to extend the clustering framework, deepen the understanding of urban form and mobility patterns, and enhance the practical applicability of cross-city comparisons.

Polycentricity

This research considered the spatial distribution of population, shops, and offices within each urban area, providing an indication of how inhabitants and activities are dispersed across the city. However, a more structural perspective could be gained by explicitly analyzing the degree of polycentricity. As introduced in Chapter 2, several methods exist to measure polycentricity, although their applicability may vary depending on the specific research objective and context.

For future research, systematically defining and measuring polycentricity could provide deeper insights into the relationship between urban form and travel patterns. Polycentric cities may support shorter trip lengths, more decentralized mobility patterns, and different modal choices compared to monocentric cities. Incorporating one or more polycentricity indicators, adapted to the scale and characteristics of each city, could therefore strengthen the understanding of how spatial organization influences transportation dynamics.

Tailored Solutions

This research identified distinct clusters of cities, providing a foundation for cross-city learning. A logical next step would be to focus on identifying specific strategies and interventions that have proven effective within these groups of cities. Developing an overview of strategies, implementation experiences, and the factors that influence success or failure within each cluster would significantly support decision-making processes. While the clustering results highlight similarities, it remains essential to recognize that solutions may still need to be adapted to the specific contexts of individual cities. A more tailored understanding would help cities not only to learn from comparable cases, but also to anticipate necessary adjustments when applying strategies to their own urban context.

Interrelation between Indicators

A final recommendation for future research is to examine the relationships between specific indicators in more detail, building on the significant correlations presented in Figure 4.6. For example, the negative correlation between congestion levels and the 95th percentile of population density, or the positive correlation with the public transport modal share, could be explored further. While it remains important to emphasize that correlation does not imply causation, a more detailed analysis could uncover underlying urban dynamics that are currently not clear. Alternatively, it could confirm that certain observed correlations are coincidental rather than causal, contributing to a more complete understanding of the results.

References

Badhruudeen, M., Derrible, S., Verma, T., Kermanshah, A., & Furno, A. (2022). A geometric classification of world urban road networks. *Urban Science*, *6*(1), 11. https://doi.org/10.3390/urbansci6010011

Bellock, K. E. (2021). Alphashape: Concave hulls in python [Python package].

Bergelings, E., & Marchetti, E. (2024, August). *Social aspects of low emission zones: Sofia case study* (Prepared as part of a study by the Institute for European Environmental Policy for the Clean Air Fund). Institute for European Environmental Policy (IEEP).

Boeing, G. (2024). Modeling and analyzing urban networks and amenities with osmnx [Python package]. https://geoffboeing.com/publications/osmnx-paper/

Borruso, G. (2003). Network density and the delimitation of urban areas. *Transactions in GIS*, *7*(2), 177–191.

Braess, D. (1968). Über ein paradoxon aus der verkehrsplanung [In German]. *Unternehmensforschung*, *12*, 258–268.

Buehler, R., Teoman, D. C., & Shelton, B. (2021). Promoting bicycling in car-oriented cities: Lessons from washington, dc and frankfurt am main, germany. *Urban Science*, *5*(3). https://doi.org/10.3390/urbansci5030058

Cardillo, A., Scellato, S., Latora, V., & Porta, S. (2006). Structural properties of planar graphs of urban street patterns. *Physical Review E*, *73*(6), 066107. https://doi.org/10.1103/PhysRevE.73.066107

Centraal Bureau voor de Statistiek. (2023). Stedelijkheid (van een gebied). https://www.cbs.nl/nl-nl/onze-diensten/methoden/begrippen/stedelijkheid--van-een-gebied--

Chen, Y., Wang, J., Long, Y., Zhang, X., & Li, X. (2022). Defining urban boundaries by characteristic scales. *Computers, Environment and Urban Systems*, *94*, 101799. https://doi.org/10.1016/j.compenvurbsys.2022.101799

Coenegrachts, E., Vanelslander, T., Verhetsel, A., & Beckers, J. (2024). Analyzing shared mobility markets in europe: A comparative analysis of shared mobility schemes across 311 european cities. *Journal of Transport Geography*, *118*, 103918. https://doi.org/10.1016/j.jtrangeo.2024.103918

Costa, L. d. F., & Tokuda, E. K. (2022). A similarity approach to cities and features. *The European Physical Journal B*, *95*, 155. https://doi.org/10.1140/epjb/s10051-022-00420-y

Crucitti, P., Latora, V., & Porta, S. (2006). Centrality in networks of urban streets. *Chaos*, *16*(1), 015113. https://doi.org/10.1063/1.2150162

da Costa, N. (2024, July). The end of the car city in portugal. In R. Lois-González & J. R. Fernandes (Eds.), *Urban change in the iberian peninsula*. Springer, Cham. https://doi.org/10.1007/978-3-031-59679-7_17

Dijkstra, L., & Poelman, H. (2012). *Cities in europe: The new oecd-ec definition* (tech. rep. No. RF 01/2012). Directorate-General for Regional and Urban Policy, European Commission. https://circabc.europa.eu/w/browse/59bfa33a-8f4b-413d-8552-ac0a93ad7e5f

Ding, R., Ujang, N., Hamid, H., & Manan, M. S. A. (2019). Application of complex networks theory in urban traffic network researches. *Networks and Spatial Economics*, *19*(4), 1123–1145. https://doi.org/10.1007/s11067-019-09455-8

EMTA, E. M. T. A. .-. (2022). Emta barometer 2022: 16th edition, based on 2020 data. https://www.emta.com

EMTA, E. M. T. A. .-. (2024). 2024 emta barometer on public transport in metropolitan areas [Based on data from 2023, with specific references to part 3 in this document.]. https://www.emta.com

Eurostat. (2022). Regions in europe — 2022 interactive edition.

Eurostat. (2023). Glossary: Degree of urbanisation (degurba). https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Degree_of_urbanisation

Feng, Y., Wang, H., Chang, C., & Lu, H. (2022). Intrinsic correlation with betweenness centrality and distribution of shortest paths. *Mathematics*, *10*(2521), 1–18. https://doi.org/10.3390/math10142521

Fu, Z., Zhou, K., & Liang, F. (2017). Identifying urban subcenters from commuting fluxes: A case study of wuhan, china. *IEEE Access*, *5*, 10161–10171. https://doi.org/10.1109/ACCESS.2017.2706485

Gehl, J. (2010). *Cities for people*. Island Press.

Greenacre, M., Groenen, P. J. F., Hastie, T., Iodice D'Enza, A., Markos, A., & Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, *2*, 100. https://doi.org/10.1038/s43586-022-00184-w

Kopczewska, K., Kubara, M., & Kopyt, M. (2024). Population density as the attractor of business to the place. *Scientific Reports*, *14*(1), 22234. https://doi.org/10.1038/s41598-024-73341-8

Lemoy, R. (2024). Monocentric or polycentric cities? an empirical perspective. *Urban Studies*, *60*(8), 1624–1642. https://doi.org/10.1177/00420980221148792

Lin, J., & Ban, Y. (2013). Complex network topology of transportation systems. *Transport Reviews*, *33*(6), 658–685. https://doi.org/10.1080/01441647.2013.848955

Lin, J., & Ban, Y. (2017). Comparative analysis on topological structures of urban street networks. *ISPRS International Journal of Geo-Information*, *6*(10), 295. https://doi.org/10.3390/ijgi6100295

Loo, B. P. Y. (2009). Transport, urban. In R. Kitchin & N. Thrift (Eds.), *International encyclopedia of human geography* (pp. 465–469). Elsevier. https://doi.org/10.1016/B978-008044910-4.01039-7

Louf, R., & Barthelemy, M. (2014). A typology of street patterns. *Journal of the Royal Society Interface*, *11*, 20140924. https://doi.org/10.1098/rsif.2014.0924

Mapbox. (2025). Geojson.io: A simple viewer and editor for geojson data. https://geojson.io/#map=3.76/49.03/8.96

Mela, A. (2014). Urban areas. In A. C. Michalos (Ed.), *Encyclopedia of quality of life and well-being research*. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_3122

Murtagh, F., & Legendre, P. (2011). Ward's hierarchical clustering method: Clustering criterion and agglomerative algorithm. *arXiv*. https://arxiv.org/abs/1111.6285

OpenStreetMap contributors. (2025). OpenStreetMap.

Ortman, S. G., Lobo, J., & Smith, M. E. (2020). Cities: Complexity, theory and history. *PLOS ONE*, *15*(12), e0243621. https://doi.org/10.1371/journal.pone.0243621

Porta, S., Crucitti, P., & Latora, V. (2006). The network analysis of urban streets: A primal approach. *Environment and Planning B: Planning and Design*, *33*(5), 705–725. https://doi.org/10.1068/b32045

Pu, H., Li, Y., & Ma, C. (2022). Topology analysis of lanzhou public transport network based on double-layer complex network theory. *Physica A*, *592*, 126694. https://doi.org/10.1016/j.physa.2021.126694

Puissant, A., & Eick, C. F. (2024). Analyzing the composition of cities using spatial clustering. *Proceedings of UrbComp 2013*, 1–12.

Rabl, A., & De Nazelle, A. (2012). Benefits of shift from car to active transport. *Transport Policy*, *19*(1), 121–131. https://doi.org/10.1016/j.tranpol.2011.09.008

Reid, C. (2022). It's been 100 years since cars drove pedestrians off the roads. *Forbes*. https://www.forbes.com/sites/carltonreid/2022/11/08/happy-birthday-car-dependency-100-years-old-this-week/

Rodrigue, J.-P., & Ducruet, C. (2016). *The geography of transport systems* (Sixth). Routledge. https://doi.org/10.4324/9781003343196

Rodrigue, J. (2013). Urban transportation and land use. In *Handbook of transport geography and spatial systems*. Texas AI&M University at Galveston. https://doi.org/10.14315/9781464627655.n7

Samburu, P., Owino, F., & Mogoria, N. (2023). Urban spatial structure and the density of retail shops in cities. *Architectural Research*. https://www.researchgate.net/publication/375714838_Urban_Spatial_Structure_and_the_Density_of_Retail_Shops_in_Cities

Schiavina, M., Freire, S., Carioli, A., & MacManus, K. (2023). Ghs-pop r2023a – ghs population grid (r2023). https://doi.org/10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE

Scruggs, G. (2020, February). *There are 10,000 cities on planet earth. half didn't exist 40 years ago*. https://nextcity.org/urbanist-news/there-are-10000-cities-on-planet-earth-half-didnt-exist-40-years-ago

Staudt, C. L., Sazonovs, A., & Meyerhenke, H. (2016). Networkit: A tool suite for large-scale complex network analysis. *Network Science*. https://doi.org/10.1017/nws.2016.20

Strano, E., Nicosia, V., Latora, V., Porta, S., & Barthelemy, M. (2013). Elementary processes governing the evolution of road networks. *Scientific Reports*, *3*, 1–7. https://doi.org/10.1038/srep03507

Sun, B., Ting, T. U., Shi, W., & Guo, Y. (2013). Test on the performance of polycentric spatial structure as a measure of congestion reduction in megacities: The case study of shanghai. *Urban Planning Forum*, *2*, 17–23.

Taplin, M. (1998). The history of tramways and evolution of light rail. *Light Rail Transit Association*. http://lrta.info/archive/mrthistory.html

Technical University of Munich. (2024). *European mobility venture: Sustainable transport solutions in amsterdam, copenhagen, and oslo* (tech. rep.) (Report from the Munich Cluster for the Future of Mobility (MCube)). Technical University of Munich. https://www.eumove.eu

Temeljotov Salaj, A., & Lindkvist, C. M. (2020). Urban facility management. *Facilities*, *38*(11/12), 525–537. https://doi.org/10.1108/F-04-2020-0042

TomTom International BV. (2024a). About: Tomtom traffic index. https://www.tomtom.com/traffic-index/about/

TomTom International BV. (2024b). Tomtom traffic index: Ranking 2024 [Metro area data used]. https://www.https://www.tomtom.com/traffic-index/ranking/

Tsiotas, D., & Polyzos, S. (2015). Introducing a new centrality measure from network analysis for regional planning and rural development: Spatial efficiency centrality. *Annals of Operations Research*, *227*, 93–127. https://doi.org/10.1007/s10479-014-1714-4

Tsiotas, D., & Polyzos, S. (2017). The topology of urban road networks and its role to urban mobility. *Transportation Research Procedia*, *24*, 482–490. https://doi.org/10.1016/j.trpro.2017.05.087

Tundulyasaree, K. (2019, September). *Topological characterizing and clustering of public transport networks* [Doctoral dissertation, Delft University of Technology]. http://repository.tudelft.nl/

Um, J., Son, S.-W., Lee, S.-I., Jeong, H., & Kim, B. J. (2009). Scaling laws between population and facility densities. *Proceedings of the National Academy of Sciences (PNAS)*, *106*(34), 14236–14240. https://doi.org/10.1073/pnas.0901898106

UN-Habitat. (2020). What is a city? https://unhabitat.org/sites/default/files/2020/06/city_definition_what_is_a_city.pdf

United Nations. (2020). *World social report 2020: Inequality in a rapidly changing world*. https://www.un.org/development/desa/dspd/world-social-report/2020-2.html

United Nations Department of Economic and Social Affairs. (2012). Goal 11: Make cities and human settlements inclusive, safe, resilient and sustainable. https://sdgs.un.org/goals/goal11

Veneri, P. (2014). Assessing polycentric urban systems in the oecd: Country, regional and metropolitan perspectives. *European Planning Studies*, *23*(6), 1128–1145. https://doi.org/10.1080/09654313.2014.905002

Weeks, J. R. (2010). Defining urban areas. In *Remote sensing and gis applications for urban environments* (pp. 33–45). Springer. https://doi.org/10.1007/978-1-4020-4385-7_3

Wilson, R. J. (1996). *Introduction to graph theory* (Fourth). Addison Wesley Longman Limited.

World Bank. (2023). *Urban development overview*. Retrieved 2024, from https://www.worldbank.org/en/topic/urbandevelopment/overview

World Economic Forum. (2018, October). *5 big challenges facing big cities of the future*. https://www.weforum.org/stories/2018/10/the-5-biggest-challenges-cities-will-face-in-the-future/

Yamaoka, K., Kumakoshi, Y., & Yoshimura, Y. (2021). Local betweenness centrality analysis of 30 european cities. *arXiv preprint*, *arXiv:2103.11437*. https://doi.org/10.48550/arXiv.2103.11437

A

Complete data set

Table A.1: Road topology and population indicators for 32 European cities.

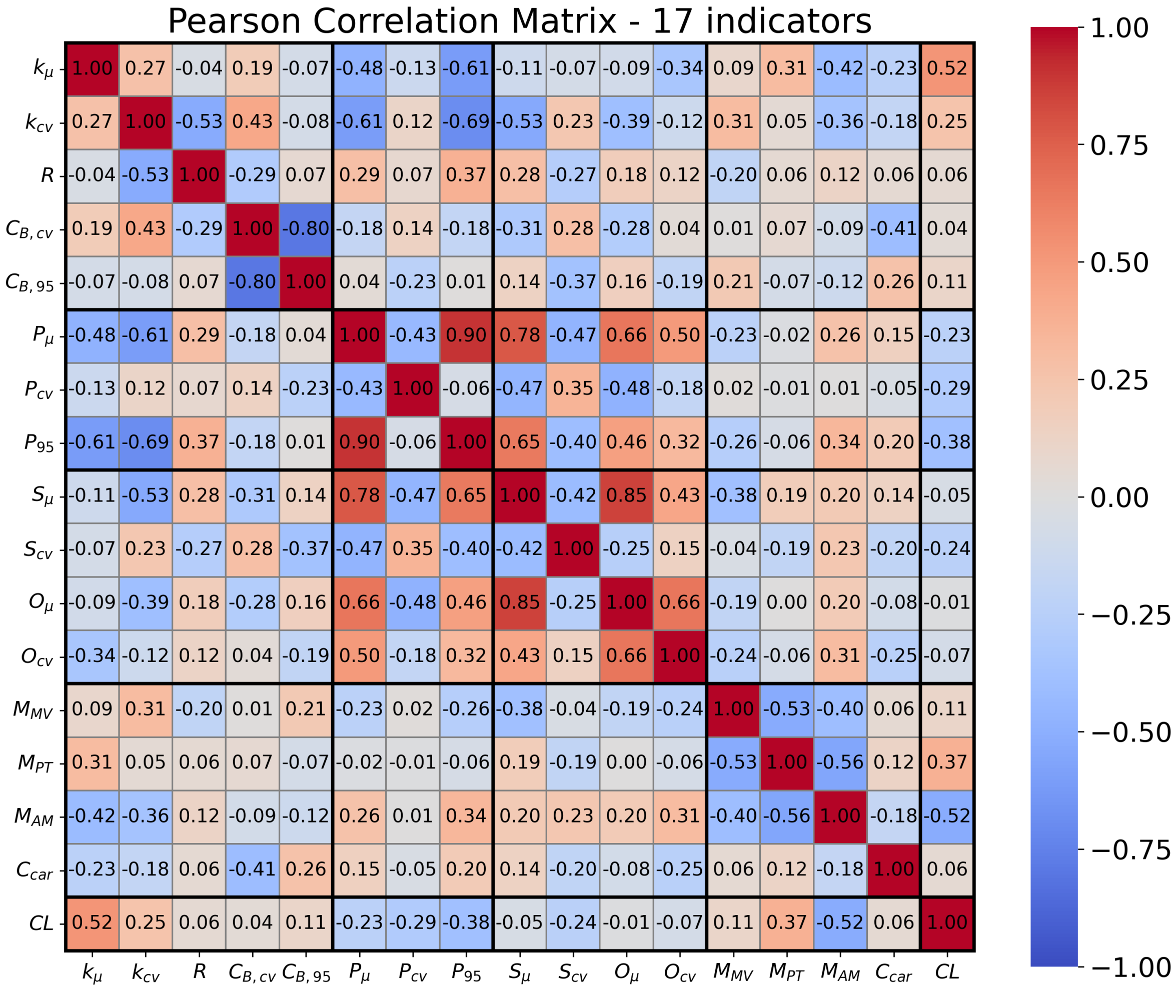
City	k_μ	k_{cv}	R	B_{cv}	B_{95}	P_μ	P_{cv}	P_{95}
Amsterdam	4.622	0.358	0.430	2.821	0.018	4,365.0	1.261	16,056.6
Barcelona	3.760	0.301	0.471	2.723	0.016	20,744.7	0.791	49,940.9
Berlin	5.121	0.330	0.407	2.547	0.024	7,052.7	0.832	18,872.0
Bilbao	3.895	0.350	0.446	2.691	0.026	5,601.6	1.931	31,407.1
Birmingham	4.441	0.426	0.375	5.107	0.006	3,857.7	1.058	10,921.3
Brussels	4.422	0.342	0.450	3.041	0.014	5,170.4	1.367	21,154.3
Bucharest	4.578	0.352	0.443	3.256	0.021	7,628.7	1.142	25,807.3
Budapest	5.264	0.338	0.474	3.745	0.010	2,859.1	1.481	10,465.6
Copenhagen	4.767	0.394	0.416	3.593	0.013	4,136.9	1.460	15,996.3
Frankfurt am Main	4.269	0.364	0.431	2.252	0.030	4,502.0	1.246	16,132.7
Helsinki	4.229	0.409	0.379	3.281	0.017	2,511.3	1.553	9,328.9
Krakow	4.391	0.405	0.411	2.455	0.029	2,881.0	1.427	11,586.5
Lisbon	4.319	0.377	0.432	4.049	0.009	4,939.8	1.899	25,043.7
London	4.643	0.397	0.417	4.897	0.005	6,371.7	1.179	20,385.5
Lyon	4.322	0.379	0.430	3.225	0.013	3,717.2	1.483	15,860.2
Madrid	3.961	0.330	0.422	3.136	0.013	14,281.8	0.925	38,959.2
Manchester	4.594	0.426	0.428	4.461	0.007	3,764.2	1.200	11,611.0
Oslo	4.447	0.403	0.358	3.751	0.021	3,901.4	1.027	11,723.5
Palma de Mallorca	4.057	0.338	0.430	2.449	0.034	7,362.9	1.166	25,441.8
Paris	4.335	0.367	0.403	5.122	0.003	6,412.5	1.312	23,431.1
Prague	4.507	0.379	0.373	3.146	0.021	4,879.2	1.100	16,009.2
Rotterdam	4.566	0.370	0.472	3.507	0.014	3,415.2	1.052	10,386.7
Sofia	4.878	0.367	0.453	2.708	0.021	4,407.6	1.202	15,144.8
Stockholm	4.628	0.408	0.422	4.492	0.012	4,812.9	1.430	17,317.9
Stuttgart	4.649	0.387	0.427	2.194	0.056	7,401.3	0.929	20,836.1
Thessaloniki	4.741	0.337	0.428	3.596	0.019	6,379.7	1.328	24,227.2
Toulouse	4.377	0.356	0.398	3.069	0.019	2,725.8	1.147	9,626.9
Turin	4.253	0.355	0.423	2.939	0.017	4,968.6	1.913	25,957.1
Valencia	3.766	0.312	0.477	2.858	0.021	8,090.4	1.391	32,293.5
Vienna	4.456	0.354	0.452	3.017	0.018	7,142.7	1.022	22,902.1
Vilnius	4.452	0.384	0.447	1.572	0.058	5,185.5	1.033	15,509.0
Warsaw	4.666	0.388	0.470	3.483	0.013	4,099.3	1.328	15,817.3

Table A.2: Economic activity, mobility and congestion indicators for 32 European cities.

City	S_μ	S_{cv}	O_μ	O_{cv}	M_{MV}	M_{PT}	M_{AM}	C_{car}	CL
Amsterdam	28.4	2.970	6.5	1.354	33	10	57	632	24
Barcelona	76.6	1.849	19.9	6.251	26	23	51	453	22
Berlin	53.3	1.772	15.6	1.386	44	19	37	431	29
Bilbao	32.6	3.323	6.7	1.747	12	22	66	440	13
Birmingham	17.5	2.978	5.9	2.200	58	7	35	391	33
Brussels	27.8	3.156	9.5	2.503	35	26	37	264	33
Bucharest	27.3	2.042	6.2	1.183	42	38	19	725	46
Budapest	21.0	2.730	6.2	1.557	35	47	18	410	32
Copenhagen	21.1	2.658	6.0	1.299	51	8	41	417	21
Frankfurt am Main	25.4	2.731	7.1	1.290	55	12	33	470	28
Helsinki	15.6	3.717	8.4	3.936	37	21	41	497	25
Krakow	26.6	2.604	6.4	1.228	41	44	14	714	36
Lisbon	24.1	2.688	6.8	1.706	60	16	24	544	23
London	28.1	2.402	7.6	1.724	35	37	28	544	33
Lyon	25.6	3.842	9.3	2.180	44	20	36	598	23
Madrid	48.9	2.313	9.1	1.382	40	24	35	817	18
Manchester	15.0	3.187	6.1	2.672	58	8	32	388	32
Oslo	14.5	2.712	5.4	1.174	44	26	29	476	21
Palma de Mallorca	26.6	2.146	5.9	0.976	55	10	35	628	19
Paris	36.9	3.031	9.9	1.880	35	22	43	368	29
Prague	29.6	2.473	6.0	1.293	23	46	31	655	31
Rotterdam	15.6	2.949	5.6	1.573	40	3	52	416	30
Sofia	50.5	2.960	9.7	2.017	36	41	23	663	35
Stockholm	22.2	2.765	5.9	1.254	39	47	13	267	20
Stuttgart	43.6	2.341	12.7	1.573	59	11	30	600	29
Thessaloniki	29.9	2.772	5.5	1.077	39	24	37	471	29
Toulouse	20.8	4.613	8.2	2.233	50	14	34	609	24
Turin	18.1	2.311	5.4	1.005	63	6	31	682	22
Valencia	20.9	2.631	5.9	1.450	41	14	43	690	19
Vienna	46.1	2.232	9.1	1.761	27	27	46	521	25
Vilnius	25.6	1.366	8.3	1.421	53	26	20	548	38
Warsaw	33.0	2.085	8.3	1.670	35	40	25	716	32

B

Correlation Matrix



	k_μ	k_{cv}	R	$C_{B,cv}$	$C_{B,95}$	P_μ	P_{cv}	P_{95}	S_μ	S_{cv}	O_μ	O_{cv}	M_{MV}	M_{PT}	M_{AM}	C_{car}	CL
k_μ	1.00	0.27	-0.04	0.19	-0.07	-0.48	-0.13	-0.61	-0.11	-0.07	-0.09	-0.34	0.09	0.31	-0.42	-0.23	0.52
k_{cv}	0.27	1.00	-0.53	0.43	-0.08	-0.61	0.12	-0.69	-0.53	0.23	-0.39	-0.12	0.31	0.05	-0.36	-0.18	0.25
R	-0.04	-0.53	1.00	-0.29	0.07	0.29	0.07	0.37	0.28	-0.27	0.18	0.12	-0.20	0.06	0.12	0.06	0.06
$C_{B,cv}$	0.19	0.43	-0.29	1.00	-0.80	-0.18	0.14	-0.18	-0.31	0.28	-0.28	0.04	0.01	0.07	-0.09	-0.41	0.04
$C_{B,95}$	-0.07	-0.08	0.07	-0.80	1.00	0.04	-0.23	0.01	0.14	-0.37	0.16	-0.19	0.21	-0.07	-0.12	0.26	0.11
P_μ	-0.48	-0.61	0.29	-0.18	0.04	1.00	-0.43	0.90	0.78	-0.47	0.66	0.50	-0.23	-0.02	0.26	0.15	-0.23
P_{cv}	-0.13	0.12	0.07	0.14	-0.23	-0.43	1.00	-0.06	-0.47	0.35	-0.48	-0.18	0.02	-0.01	0.01	-0.05	-0.29
P_{95}	-0.61	-0.69	0.37	-0.18	0.01	0.90	-0.06	1.00	0.65	-0.40	0.46	0.32	-0.26	-0.06	0.34	0.20	-0.38
S_μ	-0.11	-0.53	0.28	-0.31	0.14	0.78	-0.47	0.65	1.00	-0.42	0.85	0.43	-0.38	0.19	0.20	0.14	-0.05
S_{cv}	-0.07	0.23	-0.27	0.28	-0.37	-0.47	0.35	-0.40	-0.42	1.00	-0.25	0.15	-0.04	-0.19	0.23	-0.20	-0.24
O_μ	-0.09	-0.39	0.18	-0.28	0.16	0.66	-0.48	0.46	0.85	-0.25	1.00	0.66	-0.19	0.00	0.20	-0.08	-0.01
O_{cv}	-0.34	-0.12	0.12	0.04	-0.19	0.50	-0.18	0.32	0.43	0.15	0.66	1.00	-0.24	-0.06	0.31	-0.25	-0.07
M_{MV}	0.09	0.31	-0.20	0.01	0.21	-0.23	0.02	-0.26	-0.38	-0.04	-0.19	-0.24	1.00	-0.53	-0.40	0.06	0.11
M_{PT}	0.31	0.05	0.06	0.07	-0.07	-0.02	-0.01	-0.06	0.19	-0.19	0.00	-0.06	-0.53	1.00	-0.56	0.12	0.37
M_{AM}	-0.42	-0.36	0.12	-0.09	-0.12	0.26	0.01	0.34	0.20	0.23	0.20	0.31	-0.40	-0.56	1.00	-0.18	-0.52
C_{car}	-0.23	-0.18	0.06	-0.41	0.26	0.15	-0.05	0.20	0.14	-0.20	-0.08	-0.25	0.06	0.12	-0.18	1.00	0.06
CL	0.52	0.25	0.06	0.04	0.11	-0.23	-0.29	-0.38	-0.05	-0.24	-0.01	-0.07	0.11	0.37	-0.52	0.06	1.00

Figure B.1: Pearson correlation matrix showing correlations between all 17 indicators.

C
Road Length Analysis

For each city, the road network was retrieved using the defined boundary polygons and OpenStreetMap data with the OSMnx package, focusing on all public roads. The objective was to analyze, for each road type, the average length of road segments. This was computed by dividing the total length of each road category by the number of individual links in the network. In cases where a road segment carried multiple tags, it contributed to the statistics of each relevant road type. These values were first calculated on the original, unconsolidated road network.

The same procedure was then applied to a version of the network in which nodes were consolidated using a 25-meter tolerance. The comparison between unconsolidated and consolidated results is shown in Table C.1.

Table C.1: Average segment length per road type before and after node consolidation.

Road Type	**Unconsolidated** Average Segmented Length (m)	**Consolidated** Average Segmented Length (m)
motorway	947.5	976.5
trunk	446.8	538.0
primary	133.6	240.8
secondary	109.6	205.8
tertiary	104.0	185.2
residential	104.5	157.9
living street	88.2	157.7

To maintain clarity, very rare road types were excluded from this analysis. These include: `link`, `busway`, `unclassified`, `rest_area`, `escape`, `ladder`, `road`, `crossing`, `disused`, `emergency_bay`, `bus_bay`, and `destroyed`. These categories are either very small, only available for few cities or outside the main road hierarchy considered in this analysis.

D

Two & Five Cluster Characteristics

In support of the main findings in Chapter 5, this appendix provides the full indicator profiles for the two- and five-cluster results.

D.1. Two-Cluster Characteristics

Table D.1: Mean values for all 17 indicators for the two-cluster result. Standard deviations are shown in parentheses. Bold indicators show the included indicators for the Principal Component Analysis. Lisbon and Lyon were not assigned to a cluster.

Cluster	**Southern Cluster**	**Northwestern/Eastern Cluster**
Nr. Cities	**6**	**24**
$\boldsymbol{k_{\mu}}$	**3.949 (0.188)**	**4.585 (0.246)**
k_{cv}	*0.331 (0.021)*	*0.377 (0.028)*
$\boldsymbol{R}$	**0.445 (0.024)**	**0.423 (0.032)**
B_{cv}	*2.799 (0.235)*	*3.380 (0.928)*
$\boldsymbol{B_{95}}$	**0.021 (0.008)**	**0.020 (0.013)**
P_{μ}	*10,175 (6145.913)*	*4,827.7 (1563.296)*
$\boldsymbol{P_{\mathrm{cv}}}$	**1.353 (0.486)**	**1.213 (0.187)**
$\boldsymbol{P_{95}}$	**33,999.9 (9233.348)**	**16,302.1 (4999.219)**
S_{μ}	*37.3 (22.141)*	*28.1 (11.022)*
$\boldsymbol{S_{\mathrm{cv}}}$	**2.429 (0.507)**	**2.719 (0.642)**
O_{μ}	*8.8 (5.566)*	*7.8 (2.474)*
$\boldsymbol{O_{\mathrm{cv}}}$	**2.135 (2.037)**	**1.719 (0.641)**
M_{MV}	*39.5 (18.620)*	*41.8 (9.867)*
$\boldsymbol{M_{\mathrm{PT}}}$	**16.5 (7.583)**	**25.2 (14.340)**
M_{AM}	*43.5 (13.142)*	*32.3 (11.316)*
$\boldsymbol{V_{\mathrm{own}}}$	**618.3 (146.894)**	**508.0 (134.661)**
$\boldsymbol{CL}$	**18.8 (3.312)**	**29.8 (5.942)**

- **Southern Cluster**: *Barcelona, Bilbao, Madrid, Palma de Mallorca, Turin, Valencia*
- **Northwestern/Eastern Cluster**: *Amsterdam, Berlin, Birmingham, Brussels, Bucharest, Budapest, Copenhagen, Frankfurt am Main, Helsinki, Krakow, London, Manchester, Oslo, Paris, Prague, Rotterdam, Sofia, Stockholm, Stuttgart, Thessaloniki, Toulouse, Vienna, Vilnius, Warsaw*

D.2. Five-Cluster Characteristics

Table D.2: Mean values for all 17 indicators for the five main clusters and the sixth intermediate group. Standard deviations are shown in parentheses. Bold indicators were included in the principal component analysis. The final column shows the characteristics of the group of cities that were together, but could not be consistently assigned to a single cluster across methods.

Cluster	**Southern Cluster**	**Dense Multimodal**	**PT-Oriented Congested**	**Balanced Mid-Density**	**Northwestern Cluster**	*Inc. Assig. Group*
Nr. Cities	**4**	**1**	**7**	**3**	**7**	**6**
$\boldsymbol{k_\mu}$	**3.969 (0.206)**	**3.760**	**4.704 (0.289)**	**4.432 (0.141)**	**4.741 (0.344)**	**4.569 (0.188)**
k_{cv}	*0.337 (0.020)*	*0.301*	*0.375 (0.024)*	*0.391 (0.035)*	*0.367 (0.032)*	*0.371 (0.024)*
$\boldsymbol{R}$	**0.442 (0.026)**	**0.471**	**0.434 (0.036)**	**0.408 (0.027)**	**0.427 (0.020)**	**0.422 (0.037)**
B_{cv}	*2.906 (0.185)*	*2.723*	*3.384 (0.798)*	*4.082 (0.931)*	*2.104 (0.493)*	*3.253 (0.589)*
$\boldsymbol{B_{95}}$	**0.019 (0.006)**	**0.016**	**0.017 (0.008)**	**0.011 (0.006)**	**0.046 (0.019)**	**0.019 (0.006)**
P_μ	*8,235.6 (4,250.1)*	*20,744.7*	*4,732.4 (1,758.0)*	*4,179.3 (1,386.9)*	*6,546.5 (1,191.5)*	*4,450.1 (1,020.127)*
$\boldsymbol{P_{\mathrm{cv}}}$	**1.540 (0.480)**	**0.791**	**1.266 (0.148)**	**1.295 (0.172)**	**0.931 (0.101)**	**1.229 (0.165)**
$\boldsymbol{P_{95}}$	**32,154.2 (5,332.0)**	**49,940.9**	**16,459.5 (5,241.0)**	**14,770.2 (5,818.3)**	**18,405.7 (2,694.0)**	**15,753.8 (4,840.277)**
S_μ	*30.1 (13.978)*	*76.6*	*30.9 (9.377)*	*22.3 (7.828)*	*40.8 (14.054)*	*22.5 (6.497)*
$\boldsymbol{S_{\mathrm{cv}}}$	**2.644 (0.477)**	**1.849**	**2.471 (0.332)**	**3.350 (0.629)**	**1.826 (0.490)**	**2.799 (0.130)**
O_μ	*6.8 (1.636)*	*19.9*	*7.2 (1.413)*	*7.7 (1.723)*	*12.2 (3.689)*	*6.0 (0.661)*
$\boldsymbol{O_{\mathrm{cv}}}$	**1.396 (0.305)**	**6.251**	**1.525 (0.306)**	**2.382 (0.826)**	**1.460 (0.099)**	**1.295 (0.169)**
M_{MV}	*39.0 (20.897)*	*26.0*	*35.3 (6.184)*	*44.6 (10.502)*	*52.0 (7.550)*	*43.7 (14.140)*
$\boldsymbol{M_{\mathrm{PT}}}$	**16.5 (8.226)**	**23.0**	**41.9 (3.891)**	**20.7 (13.635)**	**18.7 (7.506)**	**18.3 (9.174)**
M_{AM}	*43.8 (15.650)*	*51.0*	*22.6 (5.968)*	*33.6 (9.863)*	*29.0 (8.544)*	*41.5 (10.950)*
$\boldsymbol{V_{\mathrm{own}}}$	**657.2 (157.483)**	**453.0**	**632.4 (116.151)**	**397.7 (122.806)**	**526.3 (86.558)**	**480.3 (79.203)**
$\boldsymbol{CL}$	**18.0 (3.742)**	**22.0**	**35.0 (5.164)**	**28.0 (5.099)**	**32.0 (5.196)**	**25.5 (4.037)**

- **Southern Cluster**: *Bilbao, Madrid, Turin, Valencia*
- **Dense Multimodal**: *Barcelona*
- **PT-Oriented Congested**: *Bucharest, Budapest, Krakow, London, Prague, Sofia, Warsaw*
- **Balanced Mid-Density**: *Berlin, Stuttgart, Vilnius*
- **Northwestern Cluster**: *Birmingham, Brussels, Helsinki, Manchester, Paris, Stockholm, Toulouse*
- *Inconsistently Assigned Group*: *Amsterdam, Copenhagen, Frankfurt am Main, Oslo, Rotterdam, Thessaloniki*

E

Python Code

This chapter includes the python code that is used for the computations necessary in this study.

E.1. Polygon creation

```python
import os
import pandas as pd
import geopandas as gpd
import osmnx as ox
import matplotlib.pyplot as plt
import alphashape
from shapely.geometry import Point, MultiPolygon

# Set Fiona as the GeoPandas I/O engine
gpd.options.io_engine = "fiona"

# --- Configuration ---
threshold_distance = 200  # Expansion distance for node inclusion (meters)
alpha_value = 50.0  # Alpha parameter for shape generation
input_folder = "Initial Polygons"
epsg4326_output_folder = "EPSG4326 Polygons"
mollweide_output_folder = "Mollweide Polygons"
plot_folder = f"Polygon Plots ({threshold_distance}m)"
esri_54009_crs = "ESRI:54009"

# --- Create output folders ---
os.makedirs(epsg4326_output_folder, exist_ok=True)
os.makedirs(mollweide_output_folder, exist_ok=True)
os.makedirs(plot_folder, exist_ok=True)

# --- Process each city polygon ---
for filename in os.listdir(input_folder):
    if filename.endswith(".geojson"):
        city_name = filename.replace(".geojson", "")
        print(f"Processing city: {city_name}")

        # Load and reproject initial polygon
        polygon_path = os.path.join(input_folder, filename)
        initial_polygon = gpd.read_file(polygon_path).to_crs("EPSG:4326")

        # Retrieve road network around polygon centroid
        centroid = initial_polygon.geometry.centroid.iloc[0]
        middle_point = (centroid.y, centroid.x)
        road_network = ox.graph_from_point(
            middle_point,
            dist=40000,
            dist_type="bbox",
            network_type="drive",
            simplify=True
        )

        # Create GeoDataFrame of nodes
        nodes_gdf = gpd.GeoDataFrame(
            {
                "node": list(road_network.nodes()),
                "geometry": [Point(data["x"], data["y"])
                             for _, data in road_network.nodes(data=True)]
            },
            crs="EPSG:4326"
        )

        # Select nodes inside initial polygon
        nodes_inside = nodes_gdf[
            nodes_gdf.geometry.within(initial_polygon.geometry.union_all())
        ]
        included_nodes = set(nodes_inside["node"])

        # Convert to UTM for distance-based expansion
        utm_crs = nodes_gdf.estimate_utm_crs()
        nodes_gdf = nodes_gdf.to_crs(utm_crs)
        initial_polygon_utm = initial_polygon.to_crs(utm_crs)

        # Expand node selection outward
        nodes_to_check = nodes_gdf[nodes_gdf["node"].isin(included_nodes)].copy()
        spatial_index = nodes_gdf.sindex

        while not nodes_to_check.empty:
            current_node = nodes_to_check.iloc[0]
            nodes_to_check = nodes_to_check.iloc[1:]
            current_geom = current_node.geometry

            nearby_nodes = nodes_gdf.iloc[
                list(spatial_index.intersection(current_geom.buffer(threshold_distance).
                    bounds))
            ]
            nearby_nodes = nearby_nodes[
                nearby_nodes.geometry.distance(current_geom) <= threshold_distance
            ]

            new_nodes = nearby_nodes[
                ~nearby_nodes["node"].isin(included_nodes)
            ]
            included_nodes.update(new_nodes["node"])
            nodes_to_check = pd.concat([nodes_to_check, new_nodes])

        # Convert selected nodes back to EPSG:4326
        included_nodes_gdf = nodes_gdf[
            nodes_gdf["node"].isin(included_nodes)
        ].to_crs("EPSG:4326")
        close_node_coords = list(zip(
            included_nodes_gdf.geometry.x,
            included_nodes_gdf.geometry.y
        ))

        # Generate alpha shape around included nodes
        alpha_shape = alphashape.alphashape(close_node_coords, alpha_value)

        if isinstance(alpha_shape, Point):
            print(f"Skipping {city_name}: Alpha shape resulted in a single point.")
            continue

        if alpha_shape:
            # Save EPSG:4326 boundary
            epsg4326_filename = os.path.join(
                epsg4326_output_folder, f"{city_name}_EPSG4326_boundary.geojson"
            )
            epsg4326_gdf = gpd.GeoDataFrame(geometry=[alpha_shape], crs="EPSG:4326")
            epsg4326_gdf.to_file(epsg4326_filename, driver="GeoJSON")

            # Save Mollweide-projected boundary
            mollweide_gdf = epsg4326_gdf.to_crs(esri_54009_crs)
            mollweide_filename = os.path.join(
                mollweide_output_folder, f"{city_name}_Mollweide_boundary.geojson"
            )
            mollweide_gdf.to_file(mollweide_filename, driver="GeoJSON")

            # Plot nodes and alpha shape
            plot_filename = os.path.join(
                plot_folder, f"{city_name}_polygon.png"
            )
            fig, ax = plt.subplots(figsize=(10, 10))
            ax.scatter(*zip(*close_node_coords), s=10, label="Nodes")

            if isinstance(alpha_shape, MultiPolygon):
                for poly in alpha_shape.geoms:
                    x, y = poly.exterior.xy
                    ax.plot(x, y, linewidth=2)
            else:
                x, y = alpha_shape.exterior.xy
                ax.plot(x, y, linewidth=2)

            ax.set_xlabel("Longitude")
            ax.set_ylabel("Latitude")
            ax.set_title(f"Polygon for {city_name}")
            ax.legend()
            ax.grid()
            plt.savefig(plot_filename, dpi=300)
            plt.close()

        print(f"Finished processing city: {city_name}")

print("All cities processed successfully!")
```

E.2. Indicator calculations

Node Degree Indicators

```
import os
import numpy as np
import pandas as pd
import geopandas as gpd
import osmnx as ox
import matplotlib.pyplot as plt
from matplotlib.ticker import FuncFormatter

# --- Configuration ---
input_dir = "EPSG4326 Polygons"
output_dir = "Node Degree Results"
plot_dir = os.path.join(output_dir, "Node Degree Histograms")
stats_csv = os.path.join(output_dir, "node_degree_stats.csv")

# Create output directories
os.makedirs(output_dir, exist_ok=True)
os.makedirs(plot_dir, exist_ok=True)

# Initialize statistics CSV if needed
columns = [
    "City",
    "Average Node Degree",
    "Standard Deviation",
    "Coefficient of Variation"
]
if not os.path.exists(stats_csv):
    pd.DataFrame(columns=columns).to_csv(stats_csv, index=False)

# --- Process each city polygon ---
results = []

for filename in os.listdir(input_dir):
    if filename.endswith(".geojson"):
        city = filename.replace("_EPSG4326_boundary.geojson", "")
```

```python
        print(f"Processing {city}...")

        # Load polygon
        polygon_path = os.path.join(input_dir, filename)
        polygon = gpd.read_file(polygon_path, engine="fiona").geometry.iloc[0]

        # Retrieve road networks
        G_truncated = ox.graph_from_polygon(
            polygon,
            network_type="drive",
            simplify=True,
            retain_all=False,
            truncate_by_edge=True
        )
        G_untruncated = ox.graph_from_polygon(
            polygon,
            network_type="drive",
            simplify=True,
            retain_all=False,
            truncate_by_edge=False
        )

        # Identify truncated nodes
        truncated_nodes = set(G_truncated.nodes) - set(G_untruncated.nodes)

        # Compute node degrees
        degree_dict = dict(G_truncated.degree())
        internal_degrees = np.array([
            deg for node, deg in degree_dict.items()
            if node not in truncated_nodes
        ])

        # Compute statistics
        avg_degree = np.mean(internal_degrees) if internal_degrees.size > 0 else 0
        std_degree = np.std(internal_degrees) if internal_degrees.size > 0 else 0
        cv_degree = (std_degree / avg_degree) if avg_degree > 0 else 0

        # Plot histogram
        fig, ax = plt.subplots()
        bins = np.arange(internal_degrees.min(), internal_degrees.max() + 2) - 0.5
        ax.hist(
            internal_degrees,
            bins=bins,
            color="greenyellow",
            edgecolor="black",
            align="mid"
        )
        ax.axvline(
            avg_degree,
            color="red",
            linestyle="--",
            linewidth=3,
            label="Average Degree"
        )

        # Format tick labels
        formatter = FuncFormatter(lambda x, _: f"{int(x):,}")
        ax.yaxis.set_major_formatter(formatter)
        ax.xaxis.set_major_formatter(formatter)

        # Set labels and styling
        ax.set_title(f"Node Degree Distribution - {city}", fontsize=18)
        ax.set_xlabel("Node Degree", fontsize=16)
        ax.set_ylabel("Frequency", fontsize=16, labelpad=10)
        ax.tick_params(axis='both', labelsize=14)
        ax.set_xticks(np.arange(1, 11))
        ax.legend(fontsize=14)

        # Save plot
        plot_path = os.path.join(plot_dir, f"{city}_node_degree.png")
        plt.savefig(plot_path, dpi=300, bbox_inches="tight")
        plt.close()

        # Store results
        results.append([
            city,
            avg_degree,
            std_degree,
            cv_degree
        ])

        print(f"{city} processed successfully!")

# Append results to CSV
pd.DataFrame(results, columns=columns).to_csv(stats_csv, mode="a", header=False, index=False)

print("Processing complete! Results saved in", stats_csv)
```

Efficiency & Betweenness Indicators

```
import os
import numpy as np
import pandas as pd
import geopandas as gpd
import osmnx as ox
import networkx as nx
import networkit as nk
import matplotlib.pyplot as plt
from matplotlib.ticker import FixedLocator, FixedFormatter

# --- Configuration ---
polygon_dir = "EPSG4326 Polygons"
efficiency_output_dir = "Efficiency Ratio Results"
betweenness_output_dir = "Betweenness Results"
betweenness_plot_dir = os.path.join(betweenness_output_dir, "Betweenness Plots")

# Create output directories
os.makedirs(efficiency_output_dir, exist_ok=True)
os.makedirs(betweenness_output_dir, exist_ok=True)
os.makedirs(betweenness_plot_dir, exist_ok=True)

# Output file paths
efficiency_stats_csv = os.path.join(efficiency_output_dir, "efficiency_stats.csv")
betweenness_stats_csv = os.path.join(betweenness_output_dir, "betweenness_stats.csv")

# Initialize CSV files if needed
efficiency_columns = ["City", "Average Shortest Path", "Network Diameter", "Efficiency Ratio"]
betweenness_columns = ["City", "Mean Betweenness", "Standard Deviation", "Coefficient of Variation", "95th Percentile"]

if not os.path.exists(efficiency_stats_csv):
    pd.DataFrame(columns=efficiency_columns).to_csv(efficiency_stats_csv, index=False)

if not os.path.exists(betweenness_stats_csv):
    pd.DataFrame(columns=betweenness_columns).to_csv(betweenness_stats_csv, index=False)

# Set NetworKit to use 10 threads
nk.setNumberOfThreads(10)

# --- Process each city ---
for filename in os.listdir(polygon_dir):
    if filename.endswith(".geojson"):
        city = filename.replace("_EPSG4326_boundary.geojson", "")
        print(f"Processing {city}...")

        # Load polygon and retrieve road network
        polygon_path = os.path.join(polygon_dir, filename)
        polygon = gpd.read_file(polygon_path, engine="fiona").geometry.iloc[0]

        G_nx = ox.graph_from_polygon(
            polygon,
            network_type="drive",
            simplify=True,
            retain_all=False,
            truncate_by_edge=True
        )
        G_nx = ox.project_graph(G_nx)
        G_nx = ox.simplification.consolidate_intersections(
            G_nx, tolerance=25, rebuild_graph=True, dead_ends=True, reconnect_edges=True
        )
        G_nx = ox.project_graph(G_nx, to_crs="EPSG:4326")

        # --- Efficiency Analysis ---

        # Keep only the largest strongly connected component
        largest_scc = max(nx.strongly_connected_components(G_nx), key=len)
        G_nx_sub = G_nx.subgraph(largest_scc).copy()

        # Convert to NetworKit directed graph
        nkG = nk.nxadapter.nx2nk(G_nx_sub, weightAttr=None)

        # Compute diameter (undirected)
        nkG_undirected = nk.graph.Graph(nkG, weighted=False, directed=False)
        diameter_algo = nk.distance.Diameter(nkG_undirected, algo=nk.distance.DiameterAlgo.Exact)
        diameter_algo.run()
        network_diameter = diameter_algo.getDiameter()[0]

        # Compute average shortest path length
        total_distance, count = 0, 0

        for node in nkG.iterNodes():
            bfs = nk.distance.BFS(nkG, node, storePaths=False)
            bfs.run()
            distances = np.array(bfs.getDistances())
            valid_distances = distances[(distances > 0) & (distances != float("inf"))]
            total_distance += valid_distances.sum()
            count += valid_distances.size

        avg_shortest_path = total_distance / count if count > 0 else None
        efficiency_ratio = avg_shortest_path / network_diameter if network_diameter > 0 else None

        # Save efficiency statistics
        efficiency_row = pd.DataFrame([
            [city, avg_shortest_path, network_diameter, efficiency_ratio]
        ], columns=efficiency_columns)
        efficiency_row.to_csv(efficiency_stats_csv, mode="a", header=False, index=False)

        # --- Betweenness Analysis ---

        # Compute node betweenness centrality
        nkG_betw = nk.nxadapter.nx2nk(G_nx)
        betweenness = nk.centrality.Betweenness(nkG_betw, normalized=True, computeEdgeCentrality=False)
        betweenness.run()
        betweenness_scores = np.array(betweenness.scores())

        mean_bc = betweenness_scores.mean()
        std_bc = betweenness_scores.std()
        cv_bc = (std_bc / mean_bc) if mean_bc > 0 else 0
        perc_95_bc = np.percentile(betweenness_scores, 95)

        # Save betweenness statistics
        betweenness_row = pd.DataFrame([
            [city, mean_bc, std_bc, cv_bc, perc_95_bc]
        ], columns=betweenness_columns)
        betweenness_row.to_csv(betweenness_stats_csv, mode="a", header=False, index=False)

        # --- Betweenness Plot ---

        fig, ax = plt.subplots()
        nodes_sorted = sorted(G_nx.nodes, key=lambda x: betweenness_scores[list(G_nx.nodes).index(x)])
        node_sizes = [betweenness_scores[list(G_nx.nodes).index(node)] * 500 for node in nodes_sorted]
        node_colors = [betweenness_scores[list(G_nx.nodes).index(node)] for node in nodes_sorted]

        edges_gdf = ox.graph_to_gdfs(G_nx, nodes=False)
        edges_gdf.plot(ax=ax, linewidth=0.5, edgecolor="gray", zorder=1)

        sc = ax.scatter(
            [G_nx.nodes[node]['x'] for node in nodes_sorted],
            [G_nx.nodes[node]['y'] for node in nodes_sorted],
            s=node_sizes,
            c=node_colors,
            cmap='viridis',
            alpha=0.95,
            edgecolor='black',
            zorder=2,
            vmax=perc_95_bc
        )

        cbar = plt.colorbar(sc, ax=ax)
        cbar.set_label("Node Centrality", fontsize=16, labelpad=10)
        cbar.ax.tick_params(labelsize=14)

        ax.set_title(f"Betweenness Centrality - {city}", fontsize=18)
        ax.set_xlabel("Longitude", fontsize=14, labelpad=5)
        ax.set_ylabel("Latitude", fontsize=14, labelpad=10)

        xticks = ax.get_xticks()
        yticks = ax.get_yticks()
        ax.xaxis.set_major_locator(FixedLocator(xticks))
        ax.xaxis.set_major_formatter(FixedFormatter([f"{x:.2f}" for x in xticks]))
        ax.yaxis.set_major_locator(FixedLocator(yticks))
        ax.yaxis.set_major_formatter(FixedFormatter([f"{y:.2f}" for y in yticks]))
        ax.tick_params(axis='both', labelsize=12)

        plot_path = os.path.join(betweenness_plot_dir, f"{city}_betweenness.png")
        plt.savefig(plot_path, dpi=300, bbox_inches="tight")
        plt.close()

        print(f"{city} processed successfully!")

print("All cities processed successfully!")
```

Population Indicators

```
import os
import numpy as np
import pandas as pd
import geopandas as gpd
import rasterio
from rasterio.mask import mask
from rasterio.warp import calculate_default_transform, reproject, Resampling
from rasterio.features import geometry_mask
import matplotlib.pyplot as plt
from matplotlib.ticker import FixedLocator, FixedFormatter, FuncFormatter
from matplotlib import colormaps

# --- Configuration ---
raster_folder = "GHS_POP_Raster_Files"
geojson_folder = "Mollweide Polygons"
output_folder = "Population Results"
population_stats_csv = os.path.join(output_folder, "population_stats.csv")

# Create output directory
os.makedirs(output_folder, exist_ok=True)

```

```
# --- Functions ---

def match_tiff_files(city_boundary, raster_folder):
    """Find TIFF files that overlap the city boundary."""
    matching_files = []
    for root, _, files in os.walk(raster_folder):
        for file in files:
            if file.endswith(".tif"):
                tiff_path = os.path.join(root, file)
                with rasterio.open(tiff_path) as src:
                    bounds = src.bounds
                    cb = city_boundary.total_bounds
                    if cb[0] <= bounds[2] and cb[2] >= bounds[0] and cb[1] <= bounds[3] and cb[3] >= bounds[1]:
                        matching_files.append(tiff_path)
    return matching_files

def process_city(city_name, geojson_folder, raster_folder, output_folder):
    """Process a city polygon: extract population statistics and plot."""
    try:
        print(f"Processing {city_name}...")

        boundary_path = os.path.join(geojson_folder, f"{city_name}_Mollweide_boundary.geojson")
        if not os.path.exists(boundary_path):
            raise FileNotFoundError(f"GeoJSON file not found: {boundary_path}")

        city_boundary = gpd.read_file(boundary_path, engine="fiona")

        tiff_files = match_tiff_files(city_boundary, raster_folder)
        if not tiff_files:
            raise FileNotFoundError(f"No matching TIFF file found for {city_name}.")

        all_density_data = []
        for tiff_file in tiff_files:
            with rasterio.open(tiff_file) as src:
                masked_data, masked_transform = mask(src, city_boundary.geometry, crop=True)
                masked_data = masked_data[0]
                masked_data[masked_data == src.nodata] = np.nan
                all_density_data.append(masked_data.flatten())

        combined_dataset = np.concatenate(all_density_data)
        converted_dataset = combined_dataset * 100  # Convert to inhabitants/km²

        total_population = np.nansum(combined_dataset)
        mean_value = np.nanmean(converted_dataset)
        std_dev = np.nanstd(converted_dataset)
        cv = std_dev / mean_value if mean_value != 0 else np.nan
        percentile_95 = np.nanpercentile(converted_dataset, 95)

        # Save statistics
        stats = {
            "City": city_name,
            "Mean Population Density": mean_value,
            "Standard Deviation": std_dev,
            "Coefficient of Variation": cv,
            "Total Population": total_population,
            "95th Percentile Density": percentile_95
        }

        df = pd.DataFrame([stats])
        df.to_csv(population_stats_csv, mode="a", header=not os.path.exists(population_stats_csv), index=False)

        # --- Plot if only one TIFF matched ---
        if len(tiff_files) == 1:
            with rasterio.open(tiff_files[0]) as src:
                masked_data, masked_transform = mask(src, city_boundary.geometry, crop=True)
                masked_data = masked_data[0]
                masked_data[masked_data == src.nodata] = np.nan
                converted = masked_data * 100

                dst_crs = "EPSG:4326"
                dst_transform, width, height = calculate_default_transform(
                    src.crs, dst_crs, converted.shape[1], converted.shape[0],
                    *rasterio.transform.array_bounds(converted.shape[0], converted.shape[1], masked_transform)
                )

                reprojected = np.empty((height, width), dtype=np.float32)
                reproject(
                    source=converted,
                    destination=reprojected,
                    src_transform=masked_transform,
                    src_crs=src.crs,
                    dst_transform=dst_transform,
                    dst_crs=dst_crs,
                    resampling=Resampling.nearest
                )

                reprojected = np.ma.masked_invalid(reprojected)
                boundary_latlon = city_boundary.to_crs(dst_crs)

                mask_shape = (reprojected.shape[0], reprojected.shape[1])
                mask_geom = geometry_mask(
                    geometries=boundary_latlon.geometry,
                    transform=dst_transform,
                    invert=True,
                    out_shape=mask_shape
                )
                reprojected.mask |= ~mask_geom

                cmap = colormaps["magma"].copy()
                cmap.set_bad(color="white")

                extent = (
                    dst_transform[2],
                    dst_transform[2] + dst_transform[0] * width,
                    dst_transform[5] + dst_transform[4] * height,
                    dst_transform[5]
                )

                fig, ax = plt.subplots(figsize=(10, 8))
                cax = ax.imshow(
                    reprojected,
                    cmap=cmap,
                    extent=extent,
                    origin="upper",
                    aspect="auto",
                    vmax=percentile_95
                )

                boundary_latlon.boundary.plot(ax=ax, edgecolor="black", linewidth=2, zorder=2)

                cbar = fig.colorbar(cax, ax=ax)
                cbar.set_label("Inhabitants per km²", fontsize=18, labelpad=10)
                cbar.ax.tick_params(labelsize=14)
                cbar.ax.yaxis.set_major_formatter(FuncFormatter(lambda x, _: f"{int(x):,}"))

                ax.set_title(f"Population Density - {city_name}", fontsize=24)
                ax.set_xlabel("Longitude", fontsize=16, labelpad=5)
                ax.set_ylabel("Latitude", fontsize=16, labelpad=10)
                ax.tick_params(axis="both", labelsize=14)

                xticks = ax.get_xticks()
                yticks = ax.get_yticks()
                ax.xaxis.set_major_locator(FixedLocator(xticks))
                ax.xaxis.set_major_formatter(FixedFormatter([f"{x:.2f}" for x in xticks]))
                ax.yaxis.set_major_locator(FixedLocator(yticks))
                ax.yaxis.set_major_formatter(FixedFormatter([f"{y:.2f}" for y in yticks]))

                plot_path = os.path.join(output_folder, f"{city_name}_population_density.png
                    ")
                plt.savefig(plot_path, dpi=300, bbox_inches="tight")
                plt.close()

    except Exception as e:
        print(f"Error processing {city_name}: {e}")

# --- Process all cities ---

for filename in os.listdir(geojson_folder):
    if filename.endswith(".geojson"):
        city_name = filename.replace("_Mollweide_boundary.geojson", "")
        process_city(city_name, geojson_folder, raster_folder, output_folder)

print("All cities processed successfully!")
```

Economic Activity Indicators

```
import os
import numpy as np
import pandas as pd
import geopandas as gpd
import osmnx as ox
import matplotlib.pyplot as plt
from shapely.geometry import box
from matplotlib.ticker import FixedLocator, FixedFormatter

gpd.options.io_engine = "fiona"

# --- Configuration ---
grid_size = 500  # Grid size in meters
boundary_folder = "EPSG4326 Polygons"
results_folder = "Results Economic Activity"
output_csv = os.path.join(results_folder, "economic_activity_stats.csv")
shops_folder = os.path.join(results_folder, "Shops Plots")
offices_folder = os.path.join(results_folder, "Offices Plots")

# Create output folders
os.makedirs(results_folder, exist_ok=True)
os.makedirs(shops_folder, exist_ok=True)
os.makedirs(offices_folder, exist_ok=True)

# Initialize statistics CSV
if not os.path.exists(output_csv):
    columns = [
        "City",
        "Grid Cells",
        "Mean Shops per km²", "Std Shops", "CV Shops",
        "Mean Offices per km²", "Std Offices", "CV Offices"
    ]
    pd.DataFrame(columns=columns).to_csv(output_csv, index=False)

# --- Functions ---

def get_projected_crs(boundary):
    """Estimate appropriate UTM CRS based on boundary centroid."""
    boundary = boundary.to_crs("EPSG:4326")
    centroid = boundary.geometry.centroid.iloc[0]
    utm_zone = int((centroid.x + 180) / 6) + 1
    crs_code = f"EPSG:{32600 + utm_zone if centroid.y >= 0 else 32700 + utm_zone}"
    return crs_code

def make_grid(boundary, grid_size):
    """Create a regular grid covering the boundary."""
    minx, miny, maxx, maxy = boundary.bounds
    cols = np.arange(minx, maxx, grid_size)
    rows = np.arange(miny, maxy, grid_size)
    grid_cells = [box(x, y, x + grid_size, y + grid_size) for x in cols for y in rows]
    return gpd.GeoDataFrame(geometry=grid_cells, crs=boundary_gdf.crs)
```

```

def compute_stats(grid, column):
    """Compute mean, std deviation, and CV for a given grid column."""
    mean = grid[column].mean()
    std_dev = grid[column].std()
    cv = std_dev / mean if mean != 0 else 0
    return mean, std_dev, cv

def format_plot(ax, title):
    """Format the plot appearance."""
    ax.set_title(title, fontsize=18)
    ax.set_xlabel("Longitude", fontsize=16, labelpad=5)
    ax.set_ylabel("Latitude", fontsize=16, labelpad=10)
    ax.tick_params(axis="both", labelsize=14)
    xticks = ax.get_xticks()
    yticks = ax.get_yticks()
    ax.xaxis.set_major_locator(FixedLocator(xticks))
    ax.xaxis.set_major_formatter(FixedFormatter([f"{x:.2f}" for x in xticks]))
    ax.yaxis.set_major_locator(FixedLocator(yticks))
    ax.yaxis.set_major_formatter(FixedFormatter([f"{y:.2f}" for y in yticks]))

# --- Process each city ---

city_files = [f for f in os.listdir(boundary_folder) if f.endswith(".geojson")]

for city_file in city_files:
    city_name = city_file.replace("_EPSG4326_boundary.geojson", "")
    print(f"Processing {city_name}...")

    boundary_gdf = gpd.read_file(os.path.join(boundary_folder, city_file))
    boundary_polygon = boundary_gdf.union_all()

    # Query shops and offices
    shops_gdf = ox.features_from_polygon(boundary_polygon, tags={"shop": True})
    offices_gdf = ox.features_from_polygon(boundary_polygon, tags={"office": True})

    # Project to appropriate CRS
    projected_crs = get_projected_crs(boundary_gdf)
    boundary_gdf = boundary_gdf.to_crs(projected_crs)
    boundary_polygon = boundary_gdf.union_all()
    shops_gdf = shops_gdf.to_crs(projected_crs)
    offices_gdf = offices_gdf.to_crs(projected_crs)

    # Create and clip grid
    grid_gdf = make_grid(boundary_polygon, grid_size)
    grid_gdf = gpd.overlay(grid_gdf, boundary_gdf, how="intersection")

    # Count shops and offices per grid cell
    for gdf, count_col in [(shops_gdf, "shop_count"), (offices_gdf, "office_count")]:
        counts = gpd.sjoin(grid_gdf, gdf, how="left", predicate="contains")
        counts = counts.groupby(counts.index).size().reset_index(name=count_col)
        grid_gdf[count_col] = 0
        grid_gdf.loc[counts["index"], count_col] = counts[count_col]

    # Convert counts to per km²
    conv_factor = 1 / ((grid_size / 1000) ** 2)
    grid_gdf["shop_km2"] = grid_gdf["shop_count"] * conv_factor
    grid_gdf["office_km2"] = grid_gdf["office_count"] * conv_factor

    # Compute statistics
    mean_shops, std_shops, cv_shops = compute_stats(grid_gdf, "shop_km2")
    mean_offices, std_offices, cv_offices = compute_stats(grid_gdf, "office_km2")

    stats = {
        "City": city_name,
        "Grid Cells": len(grid_gdf),
        "Mean Shops per km²": mean_shops, "Std Shops": std_shops, "CV Shops": cv_shops,
        "Mean Offices per km²": mean_offices, "Std Offices": std_offices, "CV Offices":
            cv_offices
    }
    pd.DataFrame([stats]).to_csv(output_csv, mode="a", header=False, index=False)

    # Reproject to EPSG:4326 for plotting
    grid_gdf_4326 = grid_gdf.to_crs(epsg=4326)
    boundary_gdf_4326 = boundary_gdf.to_crs(epsg=4326)

    for count_type, cmap, folder, label, cbar_label in [
        ("shop_km2", "Reds", shops_folder, "Shop Density", "Shops per km²"),
        ("office_km2", "Blues", offices_folder, "Office Density", "Offices per km²")
    ]:
        fig, ax = plt.subplots(figsize=(10, 8))
        boundary_gdf_4326.plot(ax=ax, facecolor="none", edgecolor="black", linewidth=1)
        grid = grid_gdf_4326.plot(
            ax=ax,
            column=count_type,
            cmap=cmap,
            edgecolor="grey",
            linewidth=0.2,
            legend=True
        )
        cbar = ax.get_figure().get_axes()[-1]
        cbar.set_ylabel(cbar_label, fontsize=16, labelpad=10)
        cbar.tick_params(labelsize=14)

        format_plot(ax, f"{label} - {city_name} ({grid_size}x{grid_size}m grid)")
        fig.savefig(os.path.join(folder, f"{city_name}_{count_type}.png"), dpi=600,
            bbox_inches="tight")
        plt.close(fig)

    print(f"Finished: {city_name}")

print(f"All cities processed successfully! Results saved in '{results_folder}'")
```

E.3. Correlation and Clustering

Correlation Matrices

```
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import pearsonr, t
from sklearn.preprocessing import StandardScaler
import matplotlib.patches as patches

# --- Configuration ---
data_file = "FullDataset.xlsx"
output_folder = "Correlation Plots"
os.makedirs(output_folder, exist_ok=True)

# --- Define indicators ---
included_columns = [
    "Average ND", "ND CV", "Efficiency Ratio",
    "BC CV", "BC 95th Percentile",
    "Mean PD", "PD CV", "PD 95th Percentile",
    "Mean Shops", "CV Shops", "Mean Offices", "CV Offices",
    "MV (%)", "PT (%)", "AM (%)", "Car Ownership", "CL"
]

excluded_ind = [
    "ND CV", "BC CV", "Mean PD",
    "Mean Shops", "Mean Offices", "MV (%)", "AM (%)"
]

column_renaming = {
    "Average ND": r"$k_\mu$",
    "ND CV": r"$k_{cv}$",
    "Efficiency Ratio": r"$R$",
    "BC CV": r"$C_{B,cv}$",
    "BC 95th Percentile": r"$C_{B,95}$",
    "Mean PD": r"$P_\mu$",
    "PD CV": r"$P_{cv}$",
    "PD 95th Percentile": r"$P_{95}$",
    "Mean Shops": r"$S_\mu$",
    "CV Shops": r"$S_{cv}$",
    "Mean Offices": r"$O_\mu$",
    "CV Offices": r"$O_{cv}$",
    "MV (%)": r"$M_{MV}$",
    "PT (%)": r"$M_{PT}$",
    "AM (%)": r"$M_{AM}$",
    "Car Ownership": r"$C_{car}$",
    "CL": r"$CL$"
}

# --- Load and preprocess data ---
df = pd.read_excel(data_file).set_index("City")
df_all = df[included_columns].copy().astype(float)
df_selected = df[[col for col in included_columns if col not in excluded_ind]].astype(float)

scaler = StandardScaler()
df_scaled_all = pd.DataFrame(scaler.fit_transform(df_all), columns=df_all.columns, index=df_all.index)
df_scaled_selected = pd.DataFrame(scaler.fit_transform(df_selected), columns=df_selected.columns, index=df_selected.index)

# --- Correlation computation ---
def compute_pearson(df):
    corr = df.corr(method="pearson")
    pvals = pd.DataFrame(np.ones_like(corr), index=corr.index, columns=corr.columns)
    for i in range(len(df.columns)):
        for j in range(i + 1, len(df.columns)):
            _, p = pearsonr(df.iloc[:, i], df.iloc[:, j])
            pvals.iat[i, j] = pvals.iat[j, i] = p
    return corr, pvals

# --- Heatmap plotting function ---
def plot_correlation_heatmap(matrix, title, filename, included_subset, mask_non_significant=False, pvals=None, alpha=0.05, annot_size=10, label_fontsize=16):
    if mask_non_significant and pvals is not None:
        matrix = matrix.where(pvals < alpha)

    renamed_matrix = matrix.rename(index=column_renaming, columns=column_renaming)

    group_bounds = [
        included_columns[0:5],
        included_columns[5:8],
        included_columns[8:12],
        included_columns[12:16],
        included_columns[16:17]
    ]

    grouped_cols = []
    for group in group_bounds:
        group_kept = [column_renaming[col] for col in group if col in included_subset]
        if group_kept:
            grouped_cols.append(group_kept)

    reordered_labels = [col for group in grouped_cols for col in group]
    reordered = renamed_matrix.loc[reordered_labels, reordered_labels]
    split_indices = np.cumsum([len(group) for group in grouped_cols[:-1]])

    plt.figure(figsize=(12, 10))
    ax = sns.heatmap(
        reordered,
        vmin=-1, vmax=1, annot=True, fmt=".2f",
        annot_kws={"color": "black", "size": annot_size},
        cmap="coolwarm", linewidths=0.5, linecolor='gray',
        cbar=True, square=True
    )
    ax.collections[0].colorbar.ax.tick_params(labelsize=16)
    ax.set_xticklabels(ax.get_xticklabels(), rotation=0, fontsize=label_fontsize)
    ax.set_yticklabels(ax.get_yticklabels(), rotation=0, fontsize=label_fontsize)

    for idx in split_indices:
        ax.axhline(idx, color='black', linewidth=2)
        ax.axvline(idx, color='black', linewidth=2)

    border = patches.Rectangle(
        (0, 0), len(reordered.columns), len(reordered.columns),
        fill=False, edgecolor='black', linewidth=2,
        transform=ax.transData, clip_on=False
    )
    ax.add_patch(border)

    plt.title(title, fontsize=20)
    plt.savefig(os.path.join(output_folder, filename), bbox_inches="tight", dpi=300)
    plt.show()

# --- Run correlations ---
all_corr, all_pvals = compute_pearson(df_scaled_all)
selected_corr, _ = compute_pearson(df_scaled_selected)

n = df_scaled_all.shape[0]
df_degrees = n - 2
alpha = 0.05
r_threshold = np.sqrt((t.ppf(1 - alpha / 2, df_degrees) ** 2) / (df_degrees + t.ppf(1 - alpha
     / 2, df_degrees) ** 2))
print(f"For n={n}, minimum correlation for significance (p < {alpha}) is: {r_threshold:.3f}")

# --- Plot heatmaps ---
plot_correlation_heatmap(
    matrix=all_corr,
    pvals=all_pvals,
    mask_non_significant=True,
    title="Significant Pearson Correlations - 17 indicators (p < 0.05)",
    filename="pearson_significant_17ind.png",
    included_subset=included_columns,
    annot_size=12,
    label_fontsize=12
)

plot_correlation_heatmap(
    matrix=selected_corr,
    title="Pearson Correlation Matrix - 10 indicators",
    filename="pearson_full_10ind.png",
    included_subset=[col for col in included_columns if col not in excluded_ind],
    annot_size=16,
    label_fontsize=20
)

print("Correlation plots saved to:", output_folder)
```

Principal Component Analysis

```
# === IMPORTS ===
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA

# === SETTINGS ===
file_path = "FullDataset.xlsx"
output_dir = "PCA Plots"
os.makedirs(output_dir, exist_ok=True)

# --- Define indicators ---
included_columns = [
    "Average ND", "ND CV", "Efficiency Ratio",
    "BC CV", "BC 95th Percentile",
    "Mean PD", "PD CV", "PD 95th Percentile",
    "Mean Shops", "CV Shops", "Mean Offices", "CV Offices",
    "MV (%)", "PT (%)", "AM (%)", "Car Ownership", "CL"
]

excluded_ind = [
    "ND CV", "BC CV", "Mean PD",
    "Mean Shops", "Mean Offices", "MV (%)", "AM (%)"
]

column_renaming = {
    "Average ND": r"$k_\mu$",
    "ND CV": r"$k_{cv}$",
    "Efficiency Ratio": r"$R$",
    "BC CV": r"$C_{B,cv}$",
    "BC 95th Percentile": r"$C_{B,95}$",
    "Mean PD": r"$P_\mu$",
    "PD CV": r"$P_{cv}$",
    "PD 95th Percentile": r"$P_{95}$",
    "Mean Shops": r"$S_\mu$",
    "CV Shops": r"$S_{cv}$",
    "Mean Offices": r"$O_\mu$",
    "CV Offices": r"$O_{cv}$",
    "MV (%)": r"$M_{MV}$",
    "PT (%)": r"$M_{PT}$",
    "AM (%)": r"$M_{AM}$",
    "Car Ownership": r"$C_{car}$",
    "CL": r"$CL$"
}

# === LOAD AND STANDARDIZE ===
df = pd.read_excel(file_path).set_index("City")
df_selected = df[[col for col in included_columns if col not in excluded_ind]]

scaler = StandardScaler()
df_scaled = pd.DataFrame(
    scaler.fit_transform(df_selected),
    columns=df_selected.columns, index=df_selected.index
)

# === PCA ===
pca = PCA()
pca_transformed = pca.fit_transform(df_scaled)
cumulative_variance = np.cumsum(pca.explained_variance_ratio_)
num_components_75 = np.argmax(cumulative_variance >= 0.75) + 1

pca_results_df = pd.DataFrame(
    pca_transformed[:, :num_components_75],
    columns=[f"PC{i+1}" for i in range(num_components_75)],
    index=df_scaled.index
)

# === SCREE PLOT ===
def plot_scree_plot(pca_model, filename):
    explained_variance = pca_model.explained_variance_ratio_ * 100
    cumulative_variance = np.cumsum(pca_model.explained_variance_ratio_) * 100

    plt.figure(figsize=(10, 6))
    plt.plot(range(1, len(explained_variance) + 1), cumulative_variance,
             marker="D", markersize=8, linestyle="--", linewidth=3, label="Cumulative Variance")
    plt.bar(range(1, len(explained_variance) + 1), explained_variance,
            alpha=0.6, label="Explained Variance")
    plt.axhline(y=75, color="r", linestyle="--", linewidth=2, label="75% Threshold")
    plt.xlabel("Principal Components", size=16)
    plt.ylabel("Explained Variance (%)", size=16)
    plt.title("Scree Plot", size=20)
    plt.xticks(ticks=range(1, 11), size=14)
    plt.yticks(size=14)
    plt.legend()
```

```python
    plt.grid(True)
    plt.savefig(os.path.join(output_dir, filename), bbox_inches="tight", dpi=300)
    plt.show()

plot_scree_plot(pca, "scree_plot_excl7ind.png")

# === LOADINGS HEATMAP ===
def get_pca_loadings(pca_model, feature_names, n_components):
    return pd.DataFrame(
        pca_model.components_[:n_components],
        columns=feature_names,
        index=[f"PC{i+1}" for i in range(n_components)]
    )

pca_loadings_df = get_pca_loadings(pca, df_selected.columns, num_components_75)

def plot_pca_loadings_heatmap(loadings_df, filename):
    loadings_df = loadings_df.rename(columns=column_renaming)
    max_abs_val = np.abs(loadings_df.values).max()

    plt.figure(figsize=(12, 6))
    sns.heatmap(
        loadings_df,
        annot=True,
        fmt=".2f",
        annot_kws={"size": 16, "color": "black"},
        cmap="coolwarm",
        center=0,
        linewidths=0.5,
        vmin=-max_abs_val,
        vmax=max_abs_val
    )
    plt.title("PCA Component Weights", size=24)
    plt.xlabel("Indicators", size=16)
    plt.ylabel("Principal Components", size=16)
    plt.xticks(rotation=0, size=14)
    plt.yticks(rotation=0, size=14)
    plt.savefig(os.path.join(output_dir, filename), bbox_inches="tight", dpi=300)
    plt.show()

plot_pca_loadings_heatmap(pca_loadings_df, "pca_loadings_excl7ind.png")

# === TRUE LOADINGS AND CONTRIBUTIONS ===
def get_true_pca_loadings(pca_model, feature_names, n_components):
    loadings = (
        pca_model.components_[:n_components].T
        * np.sqrt(pca_model.explained_variance_[:n_components])
    )
    return pd.DataFrame(
        loadings.T,
        columns=feature_names,
        index=[f"PC{i+1}" for i in range(n_components)],
    )

true_loadings_df = get_true_pca_loadings(pca, df_selected.columns, num_components_75)
variable_contributions_df = (true_loadings_df ** 2).T * 100
variable_contributions_df.rename(index=column_renaming, inplace=True)

# === CONTRIBUTIONS HEATMAP ===
def plot_variable_contributions_with_total(variable_contributions, filename_heatmap):
    total_explained = variable_contributions.sum(axis=1).rename("Total")
    full_matrix = pd.concat([variable_contributions, total_explained], axis=1)
    mask = np.zeros_like(full_matrix, dtype=bool)
    total_col_idx = full_matrix.columns.get_loc("Total")
    mask[:, total_col_idx] = True

    plt.figure(figsize=(14, 6))
    ax = sns.heatmap(
        full_matrix,
        annot=True, fmt=".2f",
        annot_kws={"size": 14, "color": "black"},
        cmap="YlGn", linewidths=0.5, mask=mask,
        vmin=0, vmax=100, cbar=True
    )

    for y in range(full_matrix.shape[0]):
        val = full_matrix.iloc[y, total_col_idx]
        ax.text(
            total_col_idx + 0.5, y + 0.5, f"{val:.2f}",
            ha="center", va="center", fontsize=14, color="black", fontweight="bold"
        )

    ax.axvline(total_col_idx, color="black", linewidth=2)
    plt.title("Variance of Each Indicator Explained by Principal Components (%)", size=18)
    plt.xlabel("Principal Components", size=16)
    plt.ylabel("Indicators", size=16)
    plt.xticks(rotation=0, size=14)
    plt.yticks(rotation=0, size=14)
    plt.savefig(os.path.join(output_dir, filename_heatmap), bbox_inches="tight", dpi=300)
    plt.show()

plot_variable_contributions_with_total(variable_contributions_df, "
    pca_contributions_with_total_column.png")

# === FINAL STATS ===
print(f"Number of PCs to reach 75% variance: {num_components_75}")
print(f"Cumulative explained variance: {cumulative_variance[num_components_75 - 1] * 100:.2f
    }%")
print(f"Average explained variance per indicator: {variable_contributions_df.sum(axis=1).mean
    ():.2f}%")
```

K-Clustering

```
import os
os.environ["OMP_NUM_THREADS"] = "1"

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from itertools import combinations
from joblib import Parallel, delayed
from collections import defaultdict
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.cluster import KMeans
from sklearn.metrics import silhouette_score, silhouette_samples
from scipy.spatial.distance import cdist

# === SETTINGS ===
file_path = "FullDataset.xlsx"
output_dir_kmeans = os.path.join("Clustering Plots", "KMeansPlusPlus_Test")
output_dir_kmedoids = os.path.join("Clustering Plots", "KMedoids Exhaustive")
os.makedirs(output_dir_kmeans, exist_ok=True)
os.makedirs(output_dir_kmedoids, exist_ok=True)

# --- Define indicators ---
included_columns = [
    "Average ND", "ND CV", "Efficiency Ratio",
    "BC CV", "BC 95th Percentile",
    "Mean PD", "PD CV", "PD 95th Percentile",
    "Mean Shops", "CV Shops", "Mean Offices", "CV Offices",
    "MV (%)", "PT (%)", "AM (%)", "Car Ownership", "CL"
]

excluded_ind = [
    "ND CV", "BC CV", "Mean PD",
    "Mean Shops", "Mean Offices", "MV (%)", "AM (%)"
]

# === LOAD & PCA ===
df = pd.read_excel(file_path).set_index("City")
```

```
df_selected = df[[col for col in included_columns if col not in excluded_ind]].astype(float)

scaler = StandardScaler()
df_scaled = pd.DataFrame(scaler.fit_transform(df_selected), index=df_selected.index, columns=df_selected.columns)

pca = PCA()
pca_transformed = pca.fit_transform(df_scaled)
cumulative_variance = np.cumsum(pca.explained_variance_ratio_)
n_components_75 = np.argmax(cumulative_variance >= 0.75) + 1

pca_df = pd.DataFrame(pca_transformed[:, :n_components_75], index=df_scaled.index, columns=[f"PC{i+1}" for i in range(n_components_75)])

# === KMeans++ ===
k_range_kmeans = range(2, 11)
num_seeds_per_k = 1000
inertia_dict = {}
silhouette_dict = {}
best_seeds = {}
best_models = {}

for k in k_range_kmeans:
    inertias = []
    silhouettes = []
    seed_inertia = {}
    print(f"\nRunning k = {k} (KMeans++)...")
    for i in range(num_seeds_per_k):
        kmeans = KMeans(n_clusters=k, init="k-means++", n_init=50, random_state=i)
        labels = kmeans.fit_predict(pca_df)
        inertia = kmeans.inertia_
        silhouette = silhouette_score(pca_df, labels)

        inertias.append(inertia)
        silhouettes.append(silhouette)
        seed_inertia[i] = inertia

        if (i + 1) % 200 == 0:
            print(f"  Seed {i+1}/{num_seeds_per_k}")

    sorted_seeds = sorted(seed_inertia.items(), key=lambda x: x[1])
    best_seed = sorted_seeds[0][0]
    best_seeds[k] = best_seed
    inertia_dict[k] = inertias
    silhouette_dict[k] = silhouettes

    print(f" →  Best seed for k = {k}: {best_seed} (inertia = {sorted_seeds[0][1]:.2f})")

    plt.figure(figsize=(6, 4))
    plt.hist(list(seed_inertia.values()), bins=30, color="skyblue", edgecolor="black")
    plt.title(f"Inertia Distribution for k = {k}", fontsize=12)
    plt.xlabel("Inertia")
    plt.ylabel("Frequency")
    plt.tight_layout()
    plt.show()

    best_model = KMeans(n_clusters=k, init="k-means++", n_init=50, random_state=best_seed)
    best_model.fit(pca_df)
    best_models[k] = best_model

# === KMedoids Exhaustive ===
k_range_kmedoids = range(2, 9)
data_matrix = pca_df.to_numpy()
city_names = pca_df.index.tolist()
pairwise_distances = cdist(data_matrix, data_matrix, metric="euclidean")

kmedoids_inertias = []
silhouette_scores = []
label_sets = []
```

```python
def compute_inertia_and_labels(medoids_idx):
    medoids = np.array(medoids_idx)
    distances_to_medoids = pairwise_distances[:, medoids]
    closest = np.argmin(distances_to_medoids, axis=1)
    inertia = np.sum(np.min(distances_to_medoids ** 2, axis=1))
    return inertia, medoids_idx, closest


for k in k_range_kmedoids:
    print(f"Processing k={k} (KMedoids)...")
    candidates = list(combinations(range(len(city_names)), k))
    results = Parallel(n_jobs=6)(delayed(compute_inertia_and_labels)(list(m)) for m in candidates)
    best_result = min(results, key=lambda x: x[0])
    inertia, medoids_idx, labels = best_result

    kmedoids_inertias.append(inertia)
    silhouette_scores.append(silhouette_score(data_matrix, labels))
    label_sets.append(labels)

    print(f" → Best inertia (WCSS): {inertia:.2f}")

# === PLOTS ===
best_k = k_range_kmedoids[silhouette_scores.index(max(silhouette_scores))]

plt.figure(figsize=(8, 5))
plt.plot(k_range_kmedoids, kmedoids_inertias, marker="o", linestyle="--", linewidth=2)
plt.xlabel("Number of Clusters (k)", fontsize=14)
plt.ylabel("Inertia (WCSS)", fontsize=14)
plt.title("Elbow Plot - K-Medoids", fontsize=18)
plt.grid(True)
plt.tight_layout()
plt.savefig(os.path.join(output_dir_kmedoids, "kmedoids_elbow_plot.png"), dpi=300)
plt.close()

plt.figure(figsize=(8, 5))
plt.plot(k_range_kmedoids, silhouette_scores, marker="o", linestyle="--", linewidth=2)
plt.axvline(best_k, color="r", linestyle="--", linewidth=2, label=f"Best: {best_k} clusters")
plt.xlabel("Number of Clusters (k)", fontsize=14)
plt.ylabel("Silhouette Score", fontsize=14)
plt.title("Silhouette Score - K-Medoids", fontsize=18)
plt.grid(True)
plt.tight_layout()
plt.savefig(os.path.join(output_dir_kmedoids, "kmedoids_silhouette_plot.png"), dpi=300)
plt.show()

for k, labels in zip(k_range_kmedoids, label_sets):
    silhouette_vals = silhouette_samples(data_matrix, labels)
    y_lower = 10
    fig, ax = plt.subplots(figsize=(8, 6))
    for i in range(k):
        ith_vals = silhouette_vals[np.array(labels) == i]
        ith_vals.sort()
        size_cluster_i = ith_vals.shape[0]
        y_upper = y_lower + size_cluster_i
        color = plt.cm.nipy_spectral(float(i) / k)
        ax.fill_betweenx(np.arange(y_lower, y_upper), 0, ith_vals, facecolor=color, edgecolor=color, alpha=0.7)
        ax.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i + 1))
        y_lower = y_upper + 10

    avg_score = np.mean(silhouette_vals)
    ax.axvline(x=avg_score, color="red", linestyle="--", label=f"Avg = {avg_score:.2f}")
    ax.set_xlabel("Silhouette Coefficient Values", fontsize=14)
    ax.set_ylabel("Data Points", fontsize=14)
    ax.set_title(f"Silhouette Plot - K-Medoids (k = {k})", fontsize=18)
    ax.set_yticks([])
    ax.legend()
    plt.tight_layout()
    plt.savefig(os.path.join(output_dir_kmedoids, f"silhouette_barplot_k{k}.png"), dpi=300)
    plt.close()
```

```

print("\nKMeans++ and K-Medoids clustering complete. All plots and outputs saved.")
```

Ward's Method

```
# === IMPORTS ===
import os
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from scipy.cluster.hierarchy import linkage, fcluster, dendrogram
from sklearn.metrics import silhouette_score

# === SETTINGS ===
file_path = "FullDataset.xlsx"
output_dir = os.path.join("Clustering Plots", "Wards Method")
os.makedirs(output_dir, exist_ok=True)

# --- Define indicators ---
included_columns = [
    "Average ND", "ND CV", "Efficiency Ratio",
    "BC CV", "BC 95th Percentile",
    "Mean PD", "PD CV", "PD 95th Percentile",
    "Mean Shops", "CV Shops", "Mean Offices", "CV Offices",
    "MV (%)", "PT (%)", "AM (%)", "Car Ownership", "CL"
]

excluded_ind = [
    "ND CV", "BC CV", "Mean PD",
    "Mean Shops", "Mean Offices", "MV (%)", "AM (%)"
]

# === STEP 1: LOAD AND STANDARDIZE DATA ===
df = pd.read_excel(file_path).set_index("City")
df_selected = df[[col for col in included_columns if col not in excluded_ind]].astype(float)
df_full_indicators = df[included_columns].astype(float)

scaler = StandardScaler()
df_scaled = pd.DataFrame(
    scaler.fit_transform(df_selected),
    index=df_selected.index,
    columns=df_selected.columns
)

# === STEP 2: PCA ===
pca = PCA()
pca_transformed = pca.fit_transform(df_scaled)
cumulative_variance = np.cumsum(pca.explained_variance_ratio_)
n_components_75 = np.argmax(cumulative_variance >= 0.75) + 1

pca_df = pd.DataFrame(
    pca_transformed[:, :n_components_75],
    index=df_scaled.index,
    columns=[f"PC{i+1}" for i in range(n_components_75)]
)

# === STEP 3: WARD CLUSTERING ===
linkage_matrix = linkage(pca_df, method="ward")
means_rows = []
stddev_rows = []
city_rows = []
threshold_rows = []

for num_clusters in range(1, 11):
    cluster_labels = fcluster(linkage_matrix, t=num_clusters, criterion="maxclust")
    pca_clustered = pca_df.copy()
    pca_clustered["cluster"] = cluster_labels

    dendro_title = "Dendrogram (Ward's Method)" if num_clusters == 1 else f"Dendrogram (Ward's Method) - {num_clusters} Clusters"
    dendro_path = os.path.join(output_dir, f"dendrogram_ward_k{num_clusters}_clusters.png")

    if num_clusters > 1:
        max_d = linkage_matrix[-(num_clusters - 1), 2]
        min_d = linkage_matrix[-num_clusters, 2]
        threshold = (max_d + min_d) / 2
    else:
        threshold = 0

    # === PLOT DENDROGRAM WITH CITY NAMES ===
    plt.figure(figsize=(12, 8))
    plt.title(dendro_title, fontsize=30)
    dendrogram(
        linkage_matrix,
        labels=pca_clustered.index.tolist(),
        leaf_rotation=90,
        leaf_font_size=12,
        color_threshold=threshold
    )
    for collection in plt.gca().collections:
        collection.set_linewidth(3)
    if num_clusters > 1:
        plt.axhline(y=threshold, color='black', linestyle='--', linewidth=2)
    plt.ylabel("Distance", fontsize=24)
    plt.xticks(fontsize=20)
    plt.yticks(fontsize=20)
    plt.tight_layout()
    plt.savefig(
        os.path.join(output_dir, f"dendrogram_ward_k{num_clusters}_clusters_with_labels.png"),
        dpi=600
    )
    plt.close()

    # === PLOT DENDROGRAM WITHOUT CITY NAMES ===
    plt.figure(figsize=(12, 8))
    plt.title(dendro_title, fontsize=30)
    dendrogram(
        linkage_matrix,
        no_labels=True,
        color_threshold=threshold
    )
    for collection in plt.gca().collections:
        collection.set_linewidth(3)
    if num_clusters > 1:
        plt.axhline(y=threshold, color='black', linestyle='--', linewidth=2)
    plt.xlabel("Cities", fontsize=24, labelpad=10)
    plt.ylabel("Distance", fontsize=24)
    plt.xticks(fontsize=20)
    plt.yticks(fontsize=20)
    plt.tight_layout()
    plt.savefig(
        os.path.join(output_dir, f"dendrogram_ward_k{num_clusters}_clusters_no_labels.png"),
        dpi=600
    )
    plt.close()

    for cluster_label in sorted(pca_clustered["cluster"].unique()):
        cluster_indices = pca_clustered[pca_clustered["cluster"] == cluster_label].index
        cluster_data = df_full_indicators.loc[cluster_indices]

        means_rows.append({
            "Ward Clusters": num_clusters,
            "Cluster Label": cluster_label,
            "Num Cities": len(cluster_data),
            **cluster_data.mean().to_dict()
        })

        stddev_rows.append({
            "Ward Clusters": num_clusters,
            "Cluster Label": cluster_label,
            "Num Cities": len(cluster_data),
            **cluster_data.std().to_dict()
        })

        for city in cluster_indices:
            city_rows.append({
                "Ward Clusters": num_clusters,
                "Cluster Label": cluster_label,
                "City": city
            })

# === STEP 4: EXPORT TO EXCEL ===
excel_output_path = os.path.join(output_dir, "ward_cluster_outputs.xlsx")
with pd.ExcelWriter(excel_output_path, engine="xlsxwriter") as writer:
    means_df = pd.DataFrame(means_rows)
    means_df.to_excel(writer, sheet_name="Indicator_Means", index=False)

    stddev_df = pd.DataFrame(stddev_rows)
    stddev_df.to_excel(writer, sheet_name="Indicator_StdDevs", index=False)

    workbook = writer.book
    bold_format = workbook.add_format({'bold': True})
    for sheet_name, df_to_check in zip(["Indicator_Means", "Indicator_StdDevs"], [means_df, stddev_df]):
        worksheet = writer.sheets[sheet_name]
        for col_num, col_name in enumerate(df_to_check.columns):
            if col_name in df_selected.columns:
                worksheet.write(0, col_num, col_name, bold_format)

    pd.DataFrame(city_rows).to_excel(writer, sheet_name="City_Names", index=False)
    pd.DataFrame(threshold_rows).to_excel(writer, sheet_name="Cluster_Thresholds", index=False)

# === STEP 5: SILHOUETTE SCORE PLOT ===
silhouette_avg_scores = []
cluster_range = range(2, 11)

for num_clusters in cluster_range:
    cluster_labels = fcluster(linkage_matrix, t=num_clusters, criterion="maxclust")
    score = silhouette_score(pca_df, cluster_labels)
    silhouette_avg_scores.append(score)

best_score = max(silhouette_avg_scores)
best_k = cluster_range[silhouette_avg_scores.index(best_score)]
print(f"Optimal number of clusters (based on silhouette score): {best_k}")

plt.figure(figsize=(10, 6))
plt.plot(cluster_range, silhouette_avg_scores, marker='o', markersize=8, linestyle='--', linewidth=3)
plt.axvline(x=best_k, color="r", linestyle="--", linewidth=2, label=f"Best: {best_k} clusters")
plt.xlabel('Number of Clusters (k)', fontsize=14)
plt.ylabel('Silhouette Score', fontsize=14)
plt.title("Silhouette Score - Ward's Method", fontsize=18)
plt.grid(True)
# plt.legend()
plt.xticks(cluster_range)
plt.savefig(
    os.path.join(output_dir, f"silhouette_ward_{n_components_75}PCs.png"),
    dpi=300, bbox_inches="tight"
)
plt.show()
```

Cluster Evaluation

```
import os
from collections import defaultdict
import pandas as pd
```

```python
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.metrics import adjusted_rand_score
from xlsxwriter.utility import xl_rowcol_to_cell

# === Load the Excel file ===
file_path = "ClusterAssignments_3Methods_k257.xlsx"
df = pd.read_excel(file_path)

# === Create output folder for plots and Excel ===
output_folder = "Cluster Evaluation"
os.makedirs(output_folder, exist_ok=True)

# === Define cluster comparison scenarios ===
scenarios = {
    "k2": ["KMeans_k2", "KMedoids_k2", "Ward_k2"],
    "k5": ["KMeans_k5", "KMedoids_k5", "Ward_k5"],
    "k7": ["KMeans_k7", "KMedoids_k7", "Ward_k7"]
}

# === Label mapping for pretty axis labels ===
label_map = {
    "KMeans": "K-Means",
    "KMedoids": "K-Medoids",
    "Ward": "Ward"
}

# === Function to compute ARI matrix ===
def compute_ari_matrix(df, methods):
    ari_matrix = pd.DataFrame(index=methods, columns=methods, dtype=float)
    for i in methods:
        for j in methods:
            if i == j:
                ari_matrix.loc[i, j] = 1.0
            else:
                ari_matrix.loc[i, j] = adjusted_rand_score(df[i], df[j])
    return ari_matrix

# === Plotting settings ===
sns.set(style="whitegrid")
plt.rcParams.update({
    "axes.titlesize": 24,
    "axes.labelsize": 16,
    "xtick.labelsize": 20,
    "ytick.labelsize": 20
})

# === Generate and save ARI heatmaps ===
for scenario, methods in scenarios.items():
    k_val = scenario.replace("k", "")
    ari_matrix = compute_ari_matrix(df, methods)

    pretty_labels = [label_map[m.split('_')[0]] for m in methods]
    ari_matrix.index = pretty_labels
    ari_matrix.columns = pretty_labels

    plt.figure(figsize=(8, 6))
    ax = sns.heatmap(
        ari_matrix, annot=True, fmt=".2f", cmap="YlGnBu", cbar=True,
        square=True, linewidths=0.5, linecolor='gray',
        vmin=0, vmax=1.0,
        annot_kws={"fontsize": 24, "color": "white", "weight": "bold"}
    )
    plt.title(f"Adjusted Rand Index - {k_val} Clusters", fontsize=20)
    plt.xticks(rotation=45, ha='right', fontsize=18)
    plt.yticks(rotation=0, fontsize=18)

    cbar = ax.collections[0].colorbar
    cbar.ax.tick_params(labelsize=14)

    plt.tight_layout()
    plot_filename = os.path.join(output_folder, f"ARI_Heatmap_k{k_val}.png")
    plt.savefig(plot_filename)
    plt.close()

print(f"Final ARI heatmaps saved to: {output_folder}")

# === Robust inconsistency analysis and Excel export ===
jaccard_threshold = 0.5
output_excel = os.path.join(output_folder, "Inconsistent_City_Assignments_Robust.xlsx")

with pd.ExcelWriter(output_excel, engine="xlsxwriter") as writer:
    for scenario, methods in scenarios.items():
        k_val = scenario.replace("k", "")
        city_list = df["City"].tolist()

        # Step 1: Prepare cluster mappings
        clusterings = {m: df.set_index("City")[m].to_dict() for m in methods}
        city_neighbors = {city: {} for city in city_list}
        for method in methods:
            clusters = defaultdict(list)
            for city, label in clusterings[method].items():
                clusters[label].append(city)
            for city in city_list:
                city_neighbors[city][method] = set(clusters[clusterings[method][city]])

        # Step 2: Compute Jaccard similarity and store results
        jaccard_scores = []
        for city in city_list:
            sets = list(city_neighbors[city].values())
            pairwise_scores = []
            for i in range(len(sets)):
                for j in range(i + 1, len(sets)):
                    inter = len(sets[i] & sets[j])
                    union = len(sets[i] | sets[j])
                    score = inter / union if union > 0 else 0
                    pairwise_scores.append(score)
            avg_jaccard = round(sum(pairwise_scores) / len(pairwise_scores), 3)
            jaccard_scores.append(avg_jaccard)

        # Step 3: Build DataFrame
        result_df = df[["City"] + methods].copy()
        result_df["Avg_Jaccard"] = jaccard_scores

        # Step 4: Write to Excel and highlight inconsistent rows
        result_df.to_excel(writer, sheet_name=f"k{k_val}", index=False)
        workbook = writer.book
        worksheet = writer.sheets[f"k{k_val}"]
        bold_format = workbook.add_format({'bold': True})

        for row_idx, score in enumerate(jaccard_scores, start=1):  # +1 for header
            if score < jaccard_threshold:
                for col_idx in range(len(result_df.columns)):
                    cell = xl_rowcol_to_cell(row_idx, col_idx)
                    worksheet.write(cell, result_df.iloc[row_idx - 1, col_idx], bold_format)

        # Step 5: Group consistent cities into unique consensus groups
        group_dict = {}  # key: frozenset of group members, value: group_id
        group_id_counter = 1
        city_to_group = {}
        consensus_rows = []

        for idx, city in enumerate(city_list):
            score = jaccard_scores[idx]
            sets = list(city_neighbors[city].values())

            if score >= jaccard_threshold:
                # Compute intersection of all cluster neighbor sets
                common_group = set.intersection(*sets)
                group_key = frozenset(common_group)

                if group_key not in group_dict:
                    group_dict[group_key] = group_id_counter
                    group_id_counter += 1

                city_to_group[city] = group_dict[group_key]
            else:
                city_to_group[city] = None  # Mark as doubtful

        # Step 6: Prepare output DataFrame
        for city in city_list:
            group = city_to_group[city]
            if group is not None:
                group_members = [c for c, g in city_to_group.items() if g == group]
                consensus_rows.append({
                    "City": city,
                    "Consensus_Cluster": group,
                    "Group_Members": ", ".join(sorted(group_members)),
                    "Consistent": True
                })
            else:
                # Union of all cluster neighbors (doubtful group)
                all_sets = list(city_neighbors[city].values())
                merged_group = set.union(*all_sets)
                consensus_rows.append({
                    "City": city,
                    "Consensus_Cluster": "",
                    "Group_Members": ", ".join(sorted(merged_group)),
                    "Consistent": False
                })

        # Step 7: Write to Excel
        consensus_df = pd.DataFrame(consensus_rows)
        consensus_df.to_excel(writer, sheet_name=f"k{k_val}_Summary", index=False)

        # Step 8: Track group stability and detect jumping clusters
        jumping_report = []
        for method in methods:
            # Step 8.1: Group cities by their label
            label_groups = defaultdict(list)
            for city, label in df.set_index("City")[method].items():
                label_groups[label].append(city)

            for label, cities_in_group in label_groups.items():
                method_comparison = [m for m in methods if m != method]
                split_counts = []

                for other_method in method_comparison:
                    other_labels = df.set_index("City").loc[cities_in_group, other_method]
                    overlap_counts = other_labels.value_counts()
                    split_counts.append(len(overlap_counts))

                max_splits = max(split_counts)

                jumping_report.append({
                    "Reference_Method": method,
                    "Cluster_Label": label,
                    "Cities_in_Cluster": ", ".join(sorted(cities_in_group)),
                    "Max_Splits_Across_Methods": max_splits,
                    "Is_Jumping": max_splits > 1
                })

        # Step 9: Write jumping cluster analysis to Excel
        jump_df = pd.DataFrame(jumping_report)
        jump_df.to_excel(writer, sheet_name=f"k{k_val}_JumpingClusters", index=False)

print(f"Inconsistent cities with Jaccard scores saved to: {output_excel}")
```

E.4. Road Length Analysis

```
import os
import numpy as np
```

```python
import pandas as pd
import geopandas as gpd
import osmnx as ox

# --- Configuration ---
input_dir = "EPSG4326 Polygons"
output_dir = "Road Length Results"
road_length_stats_csv = os.path.join(output_dir, "road_type_stats.csv")

# Create output directory
os.makedirs(output_dir, exist_ok=True)

# Initialize statistics CSV if needed
columns = [
    "City",
    "Road Type",
    "Total Length (km)",
    "Average Segment Length (m)",
    "Number of Links",
    "Consolidated"
]
if not os.path.exists(road_length_stats_csv):
    pd.DataFrame(columns=columns).to_csv(road_length_stats_csv, index=False)

# --- Process each city polygon ---
for filename in os.listdir(input_dir):
    if filename.endswith(".geojson"):
        city = filename.replace("_EPSG4326_boundary.geojson", "")
        print(f"Processing {city}...")

        # Load polygon and retrieve road network
        polygon_path = os.path.join(input_dir, filename)
        polygon = gpd.read_file(polygon_path, engine="fiona").geometry.iloc[0]

        G_nx = ox.graph_from_polygon(
            polygon,
            network_type="drive",
            simplify=True,
            retain_all=False,
            truncate_by_edge=True
        )

        # Extract road types before consolidation
        edges = ox.graph_to_gdfs(G_nx, nodes=False)

        if "highway" in edges.columns:
            edges = edges.explode("highway")
            edges["highway"] = edges["highway"].astype(str)

            road_stats = edges.groupby("highway")["length"].agg(["sum", "mean", "count"]).reset_index()
            road_stats["sum"] /= 1000  # Convert length to kilometers
            road_stats.columns = [
                "Road Type",
                "Total Length (km)",
                "Average Segment Length (m)",
                "Number of Links"
            ]
            road_stats.insert(0, "City", city)
            road_stats.insert(5, "Consolidated", False)

            # Append statistics
            road_stats.to_csv(road_length_stats_csv, mode="a", header=False, index=False)

        # Consolidate intersections
        G_nx = ox.project_graph(G_nx)
        G_nx = ox.simplification.consolidate_intersections(
            G_nx,
            tolerance=25,
            rebuild_graph=True,
            dead_ends=True,
            reconnect_edges=True
        )
        G_nx = ox.project_graph(G_nx, to_crs="EPSG:4326")

        # Extract road types after consolidation
        consolidated_edges = ox.graph_to_gdfs(G_nx, nodes=False)

        if "highway" in consolidated_edges.columns:
            consolidated_edges = consolidated_edges.explode("highway")
            consolidated_edges["highway"] = consolidated_edges["highway"].astype(str)

            consolidated_road_stats = consolidated_edges.groupby("highway")["length"].agg(["sum", "mean", "count"]).reset_index()
            consolidated_road_stats["sum"] /= 1000
            consolidated_road_stats.columns = [
                "Road Type",
                "Total Length (km)",
                "Average Segment Length (m)",
                "Number of Links"
            ]
            consolidated_road_stats.insert(0, "City", city)
            consolidated_road_stats.insert(5, "Consolidated", True)

            # Append consolidated statistics
            consolidated_road_stats.to_csv(road_length_stats_csv, mode="a", header=False, index=False)

        print(f"{city} processed successfully!")

print("All cities processed successfully!")
```